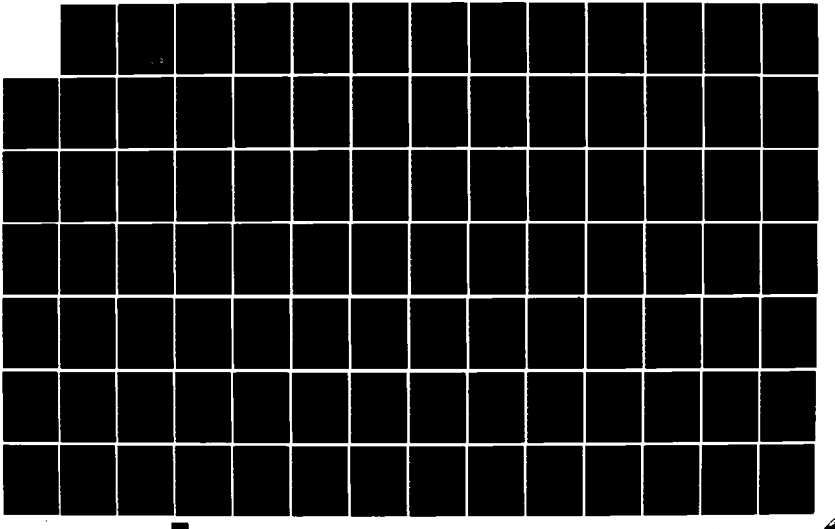
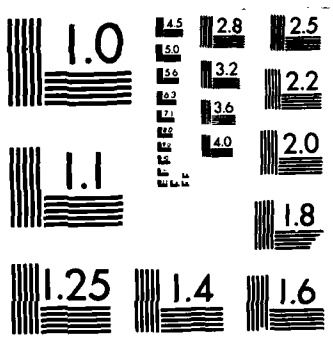
AD-A155 873 EDGE DETECTION AND GEOMETRIC METHODS IN COMPUTER VISION 1/3
(U) STANFORD UNIV CA DEPT OF COMPUTER SCIENCE
A P BLICHER FEB 85 STAN-CS-85-1041 MDA903-80-C-0102
UNCLASSIFIED F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

February 1985

Report No. STAN-CS-85-1041
Also numbered AIM-352

Edge Detection and Geometric Methods in Computer Vision

by

A. Peter Blicher

Contract MDA-903-80-C-0102
Contract N00039-82-C-0250

Department of Computer Science

Stanford University
Stanford, CA 94305

Edge Detection and Geometric Methods in Computer Vision

Edge Detection and Geometric Methods in Computer Vision

A. Peter Blicher

Abstract

Basic problems of vision are studied from the viewpoint of modern differential topology and geometry; primarily: edge detection, stereo matching, picture representation at multiple scales, and motion. Some mathematical background is provided for the non-expert.

A comprehensive review of edge detection is presented, including analyses from a mathematical perspective as well as evaluations of practical performance and promise.

Some new edge detection techniques are introduced, including a nonlinear operator based on a symmetry principle, a variational approach to global edge finding, and a least-squares localization method. A theorem is proved which shows that localizing edge position and orientation requires at least 2 orientation dependent families of convolution operators.

A unifying mathematical structure is presented for vision, notably stereo, motion stereo, optic flow (apparent flow of visual space under motion), and matching. The general matching problem is analyzed, and it is proved that generically, general matching is highly degenerate for monochrome pictures, but has a unique solution for 2 or more color dimensions. The result is extended to pictures with unknown bias and gain. Smale diagrams and level set topology are introduced as invariants useful for matching, reducing the problem to graph or tree matching. The level set topology tree is also proposed as a method of multi-scale description of the picture, and shown to be an invariant generalization of the "scale space" technique.

The motion problem is analyzed using Lie group methods, and a theorem is proved

establishing that generically 6 simultaneous values of time derivative of the monochrome picture function are necessary and sufficient to uniquely specify the 3-dimensional rigid motion of a generic given object. For 2 or more color dimensions, this is reduced to values at 3 points in the picture.

Acknowledgments

Through the period of this work, my parents have been my greatest asset, not only by their love and sacrifice, but in many, many ways.

I am grateful to my advisers, Prof. Tom Binford and Prof. Hans Bremermann, for their unflagging support and encouragement. Tom Binford generously provided the financial support and facilities which made this work possible. Hans Bremermann's kindness and helpfulness continued unabated throughout a long and rocky trek. Stephen Omohundro has helped immensely with many discussions. Lynn Hall patiently and selflessly made this effort possible for me. Dr. Harlyn Baker aided me in manifold ways. Dr. Anni Bruss, Dr. Dick Gabriel, David Marimont, Dr. David Lowe, Dr. Hans Moravec, Prof. Rod Brooks, Dr. Oussama Khatib, Allan Miller, Michael Lowry, and Sathya Narayanan have all been generous with their help, advice, and friendship. Dr. Eric Berger and Prof. Morris Hirsch gave me expert mathematical advice.

The Stanford computer science community, always generous, helpful, encouraging, and pleasant, far beyond the norm, have my heartfelt gratitude for magically transforming difficulties into pleasures. Betty Scott, Len Bosack, Ralph Gorin, Marty Frost, Marlie Yearwood, Amy Atkinson, Don Coates, Lynn Gotelli, Harry Llull, Richard Manuck, Les Earnest, and Patte Wood have all been kind far above the call of duty. Prof. John McCarthy is, if not the father, then the midwife of this environment I have enjoyed.

I am indebted to Dr. Lee Nackman of the IBM T.J. Watson Research Center for bringing the work of Koenderink and van Doorn of Utrecht to my attention.

This document was typeset using TEX, and therefore owes a debt to Prof. Donald Knuth, who with TEX has elevated the art of writing to the art of computer programming.

Arthur Keller, Dave Fuchs, and Prof. Knuth provided expert help with TEX in numerous times of need.

For leading me to this research, I must also thank the professors of engineering who taught me that the key to engineering is understanding physics; the professors of physics who taught me that the key to physics is understanding mathematics; and the professors of mathematics who taught me that mathematics has nothing to do with the material world.

This work was partially supported by ARPA contracts MDA903-80-C-0102 and N00039-82 C-0250.

Preface

I wrote this work for an audience of both vision workers and mathematicians. There are many people who could be counted in both groups, but the full intended audience has a wide spectrum of backgrounds. Among vision workers, I include students of biological vision, but it is artificial vision that I am directly addressing. It is widely appreciated that there are many common problems, but I haven't written about any specifically ***biological*** problems, such as explaining the function of some cell population. My hope is not only to communicate research results, but to convince readers in each of these fields that there is something of interest to them in the other. Many people have a bad taste from previous experience with touters of fancy mathematics in these concrete situations, with elixirs which turned out to be oversold or plain irrelevant. Mathematics is not magic; using it doesn't melt all impediments into triviality. It merely provides a structure for understanding, and an apparatus for resolving questions. *If* the questions are the right grist for the mill. Attracting these fields to each other isn't necessarily easy.

Also, it poses a problem in writing; since I have tried to keep the material accessible to the novice, some of it must necessarily be old hat to the expert, so I apologize to the expert whom I have subjected to the obvious. I have tried to make this work reasonably self-contained, including various standard definitions and results from differential topology and geometry. When these are not in the main line of thought, they have been relegated to fine print, so they can be easily skipped by those who already have the necessary background. Sometimes, standard terms are used before they are defined, and sometimes they are defined twice, partially from a lack of organization, but primarily to locate the mathematical digressions where they are most important, and avoid bogging things down where they are not. I haven't tried to be exhaustive in this, or I would have been obliged to include a complete introduction to differential topology and geometry,

something which has already been accomplished with great skill by others. The chapter Geometric Methods in Vision makes the heaviest use of abstract mathematics; therefore I have put most mathematical background material into the fine print of that chapter. Since I have assumed some of that background material in earlier chapters, the reader may find it useful to glance through it to clarify the unfamiliar, such as the implicit function theorem, or functional notation.

The 3 major chapters (A Survey of Edge Detection, Contributions to Edge Detection, and Geometric Methods in Vision) are largely independent, and can be read in any order, or in isolation. The survey has many discussions which go beyond summarizing, and should be of interest to readers who are already familiar with the literature, as well as to newcomers. The contributions chapter is probably of most interest to specialists, while the geometry chapter is likely to appeal to the more mathematically inclined.

Stanford, California
October, 1984

Table of Contents

systems. The theory for nonlinear systems is not as well developed, and not so widely known. Who can say that there is not some other isometry that is more appropriate for a nonlinear system? But how is a physiologist or psychologist to seek evidence for such an unknown object?

Finally, a third motivation for edge finding is based on the *computational* consideration of efficiency. Since boundaries are of a smaller dimension than images or regions, they are easier to handle: e.g. (for a smooth boundary, rather than a fractal) if the number of boundary points increases as $O(n)$, with $1/n$ the discretization interval (grid size), then the number of region points increases as $O(n^2)$. Also, the 1-dimensionality of a boundary provides a natural ordering for its points, which is easily translated to a processing order for a sequential algorithm. While the entirety of an image is filled with region points, only a small fraction constitute edge points.

The sparsity of edge points among image points is a major attraction of edge detection as an early step in stereoscopic vision, since it extremely diminishes the size of the search involved in matching points between the two images. Of course, this is only useful if the edge points bear some relation to things rigidly attached to fixtures in the world, so as to vindicate the assumption that edge points must match edge points. As it happens, this seems in fact to be a good assumption, and edge points appear to be more stable than more rudimentary features of intensity, such as the actual brightness values.

So far we have given a very general definition of edge detection as finding the geographical limits of a description. It is probably fair to say that few if any authors of edge detection methods thought that was what they were doing. The universal goal of edge detection algorithms is to find places in the image which a human would classify as "edges" or "boundaries." We apologize if that seems a trivial statement, but since we do not know how a person segments a scene, we are in no position to give an authoritative definition of what constitutes an "edge." Everyone agrees that a transversely translated step function

In more familiar terms, the description is nothing more than a model of the data in some system of *representation* of the pertinent knowledge. Including boundaries or regions per se in the representation language makes vacuous the notion of homogeneity. Our observations above are merely rewordings of the thesis that humans have a very rich representation language available, while machines as yet do not. Incidentally, the term *representation language* is meant to refer to the internal representation, whether it be essentially symbolic, continuous, or whatever, and shouldn't be confused with the language we use to communicate about the aspects of the representation introspectively available to us.

The problem of finding homogeneous regions can be approached either by finding the regions directly, whose difficulty increases with the complexity of description, or by finding the boundaries between the regions. Thus the first motivation for edge finding is for the purpose of segmentation into homogeneous regions in accordance with our own models of the world. It is based on *introspective* observations.

A second motivation is derived from *extrospective* observations which have been made by physiologists, perception psychologists and perceptual psychophysicists. Anatomical structures have been found which respond to abrupt changes in intensity and color as functions both of position and time. Observations have been made that indicate absolute colors and intensities, as well as their slow changes are not readily perceived, but abrupt changes are. Whether it is wise to mimic nature, or rather to attempt to mimic the precious little we think we know about nature, is problematic, despite the widespread tacit acceptance of the idea. It is worth considering that we are unlikely to find physiological processes involved with things that are not already a part of our introspective models. For example, we look for evidence that the visual system performs Fourier transforms, since Fourier transforms have a particular intuitive appeal. But they can also be viewed as only one of a myriad of possible isometries of a function space, special because they transform convolutions to multiplications. But that special property is mainly significant for *linear*

way to describe such a heterogeneous object is still to partition it into homogeneous components. The tendency is always to subdivide, perhaps reflecting the reductionist ethos widespread in modern science. One might expect that in artificial intelligence jargon, this process would be called the *chotomizing heuristic*; in fact, it is called *segmentation.* (Or borrowing from Cæsar, subdivide and conquer.)

The products of the subdivision are homogeneous entities. For a human, the homogeneity is one of description, while for the machine it is generally one of measurement. Now, a measurement is a description, and a set of descriptions is a description, so we have to explain what we mean by these terms a little more precisely. By a description, we mean something which might be quite complicated from a machine perspective, encompassing such explicit descriptions as "it gets darker and redder from right to left, with a speckling that looks like that on a trout, but which fades into a very dense network of lines in the periphery." That would not generally be found to be a homogeneous region by a machine. A measurement, on the other hand, is meant to connote something very close to the language of the transducer providing the image data, e.g. brightness or range values. Most of the definitions of homogeneity implicit in automatic segmentation programs stray little from a constancy of such a measurement, though the situation appears to be improving.

By demanding a *homogeneous* description to define a homogeneous region, we mean that the description can have no *explicit* mention of boundaries or constituent regions. The relatively weak condition of explicitness is right because an *implicit* boundary would not be a property of the description, but rather something inferred from it. Thus a description such as "the value goes linearly from 100 in the lower left to −100 in the upper right" would be judged homogeneous, despite the well defined diagonal boundary separating positive from negative values. For the time being, we are content to include as homogeneous such descriptions as "the intensity goes as a step function..."

A Survey of Edge Detection

Introduction

Why edge detection?

Edge detection, in the form of spatial differentiation, appears in the computer vision literature as early as 1955 [Dineen 1955]. This early and sustained interest arises from a perception that the types of patterns significant to a visual system consist of approximately homogeneous regions separated by abrupt boundaries. Although years of experience have shown that real digitized scenes are not easily characterized this way, the idea has persisted tenaciously, for the following reasons.

First, from an *introspective* point of view, one tends to believe the world to be composed of *objects*, each homogeneous in its cohesion, and abruptly separated from other objects and the background. That is an essential aspect of our way of perceiving the world, pervading disciplines from anatomy (where every bump, nodule, fascia, and tissue type is seen as a separate structure) to quantum mechanics (in which an essentially continuous all-pervasive field is seen to describe a separate localized particle). Whether this discretization of the world is a part of the structure of the world or of ourselves we cannot say (arguably it is impossible to say); nevertheless, it is here to stay.

Now close inspection of digital images, or for that matter, paintings from past centuries, leaves little doubt that the image of a single (realistic looking) object can but rarely be described as "homogeneous." Yet, even upon making such an observation, one's natural

---

Some mathematical background which is assumed in this chapter, such as functional notation and some results from differential topology, is explained in more detail in the fine print of the chapter Geometric Methods in Vision.

beginning steps in applying the qualitative method. The edge detection problem has not been laid to rest, alas, but I think the armamentum for its conquest, and others as well, is now closer at hand. A few dragons have been slain along the way, and their proofs are given.

in understanding how the geometry and topology interact, and I began by studying the consequences of topology. This led to results which intertwine color vision with stereo, and which clarify the role of geometric constraints in monochrome stereo vision.

The main invariant structure in these studies was the family of level sets; each level set is the set of points in the picture that all have the same brightness or color. Differential topology has studied this structure thoroughly, so we were able to say some things about how the level sets fill up the picture, and what happens when the picture is perturbed. This forms a departure point for characterizing the picture function which is independent of viewpoint. It also is related to smoothing the picture, and some other operations which people have applied to find features in pictures, e.g. zero crossings of Laplacians of Gaussians.

The next step was to study the effects of the geometry. In the spirit of modern geometry, I approached this by studying the action of a Lie group, i.e. by looking abstractly at the effect of rigid motion. This type of problem is usually easier in its differential form, i.e. for infinitesimal motions, so that was the best place to start. This put us into the business of studying motion rather than discrete views. A family of basic questions is how much can be deduced about the motion from how much data. The particular one of these questions that I studied was based on the idea that the raw data consists only of brightness or color values at fixed places on the retina, along with data on how they are changing with time. This is somewhat different than the approach many have taken in the past, where the goal has either been to find the 3-dimensional motion from the motion of points on the retina, or else to find that motion of individual retinal points itself. We were able to show that generally 6 data points of our kind are enough to specify the motion of a given surface. Again, this was related to color. The number 6 is for monochrome data; for color data, 3 points are enough.

The 2nd main topic of this thesis, geometry applied to vision, thus comprises some

suited to scrutiny under the macroscope of modern geometry. The "real" objects of vision are objects embedded in 3-dimensional space. They are subjected to various lighting conditions, and viewed in a variety of ways. It is the properties of the real, solid objects that we must deduce from the image; so we must study how the properties express themselves, and come to understand their invariances, especially invariances of qualitative features. I use "qualitative" in the sense of "qualitative dynamics," where understanding comes from topological descriptions, still quite rigorous, rather than from some formula which allows us to calculate the precise numerical values describing the state of a system. E.g., we would rather know where (and whether) an object's shape changes (say from convex to concave), than know the terms of a polynomial that specifies that shape.

The route of study then became to approach the specific, e.g. edge finding, by first understanding the overall problem. My first step was simply to write down what the spaces, maps, and groups were that were involved. This provided a structure in which to apply formal results of mathematics.

A basic fact of life in seeing 3-dimensional objects on a 2-dimensional retina is that as we change our viewpoint, or as the objects move, their 2-dimensional images undergo distortions. Understanding 3 dimensions from this 2-dimensional world must involve recognizing an object despite these distortions, and, what's more, interpreting the distortions to deduce the shape of the object and its relative motion. The exact distortion that the picture undergoes depends on the shape of the object, the motions of object and observer, and the optics that produce the picture.

Stereoscopic vision conventionally starts with 2 pictures from different views and requires finding places in the 2 pictures that correspond to a single place in the 3-dimensional scene. When this is done for enough places in the pictures, it allows triangulation to find depth (i.e. the 3rd dimension). A basic problem is to study the invariants which allow such matching to take place. Since the distortions can be complex, I was interested

rifts, ridges, dips, etc. are, and how they interlock. Then, once the image is "understood" this way, maybe to the point of hypothesizing objects, some regions may take on special importance as "edges." In this view, while 1-dimensional objects—edges—are important for representing what's in the image, they are a *result*, not a first step, of understanding. This is a somewhat heretical point of view, and it is by no means certain. But I became convinced that the understanding of local image features, e.g. labelling some features as edges, depended on getting a qualitative global understanding of the image. When I say "global" here, I don't mean that one has to understand the whole area of the picture, but rather a large enough area that the most local measurements can be put into a context. For example, from a single view through a tiny peephole one might say something about which way the shading is changing, but it takes a larger field of view to say something about the 3-dimensional object involved—whether we are looking at a convexity or concavity, a fold, an edge, an uninteresting shading gradient, or some other solid feature.

The enterprise of computer vision seeks to duplicate a feat we know from introspection, but to duplicate it by cold mathematical means. There are many styles of research, but in my thinking, this enterprise is most likely to succeed if the mathematical setting and the questions being posed are stated explicitly and precisely. Often, in fact, finding the right way to state a problem turns out to be a cornerstone of the solution. I had come to see the problem as one of describing the image appropriately on increasingly global scales, and piecing together the descriptions to arrive at local interpretations. This makes it essential to find the right ways to understand the picture function for the goal of understanding 3-dimensional structure. Differential topology and geometry seemed to be the right places to look.

Vision is teeming with geometry: the image comes from a map from a higher dimensional space with important singularities, there are natural group actions, e.g. the rigid motion group, topological invariance in the image is important, etc. These are things perfectly

through a very small peephole.

Since the very beginnings, researchers who have built vision systems which did some higher level recognition tasks have been sorely limited by the abilities of the lower level algorithms they used for input. This was a problem for ACRONYM [Brooks 1981] no less than for [Waltz 1972]. Waltz created a system which was able to understand line drawings of toy blocks. The presumption was that low-level vision could supply the line drawings. But it had turned out to be very difficult to obtain line drawings good enough to use as input, even in the blocks world domain of uniform matte surfaces in good lighting. This problem had already spawned the efforts of Binford-Horn [Horn 1972] and then, later, of [Shirai 1973]. 10 years later, [Brooks 1981] used the state-of-the-art line finder of [Nevatia and Babu 1978], and found that he had to draw inferences based on almost laughably meager low-level output. Today, reliable segmentation (dividing an image into meaningful parts) remains a paramount obstacle to image understanding.

Hence I was drawn to edge detection as a basic problem which might yield to a mathematical approach. I found that people had applied a great patchwork of techniques, but that the problem itself was very poorly understood. Edges, it seems, are a lot like obscenity, for as Mr. Justice Potter Stewart wrote of obscenity [*Jacobellis v. Ohio* 1964], he may not be able to define it, "But I know it when I see it." Everyone agrees that a perfect step function should give an edge, but there has been no adequate criterion put forth to classify any other function as edge or non-edge. There was no viable theory to bridge the gap between the local methods of the peephole and the global objects we think of as edges, if indeed the global must even first come from the local. I eventually realized that the problem of edge detection was first a problem of understanding the image intensity function, a *qualitative* understanding which must be suited to the needs of vision. In fact, I wonder if edge detection is a bit of a red herring. The best, sharpest edges are easy enough to find, all right, but it seems that the global picture may require knowing how all the local qualitative features of the image fit together: where the bumps,

anywhere within range.

Someone who has never tried to write a vision program is likely to think that it really can't be all that hard. The act of vision is so effortless for us, so transparent, that it is hard at first to imagine that anything at all must be done to bring it off, beyond the initial transduction. You digitize what's out there; the objects are delineated by their borders, which are the places where things change suddenly, so you look for the places of rapid change, figure out what the objects are, and voilà! all done. In fact, a lot of research was based on that paradigm. The illusion of simplicity seduced many people *into* thinking that it was a programming problem like many others, which could be solved by doing some intuitively obvious things, followed by bug-fixing, honing, and tuning. Unfortunately, life is not so easy.

One of the hardest things to appreciate, even to describe, is what it means to know what is in an image. In some sense, the set of all the pixel* values already has all the information about the image. But there is no ***knowledge that***, i.e. no symbolic knowledge. You can't know about the relation between any 2 pixels unless those pixels talk to each other somehow. From another perspective, knowing all the pixel values is no better than knowing them one at a time - as completely *local* information but what we really need is *global* information. And the global information needed must be exactly that information that lets us draw inferences about the physical situation that produced the image. Global information is very hard to obtain because a picture contains a lot of data--around 256K bytes† for a normal TV frame; on the order of 100,000K bytes for the human retina (for comparison, a page of text in a book is around 1K byte)--and the space of patterns to consider is of high dimension. Most people have approached this complexity problem by trying to extract information for very small regions, from a few to a few hundred pixels, and to use this information for only a few such regions at a time. This is like looking

---

* *Pixel* is a contraction of *picture element*, a single point of data in the picture.

† We take a *byte* to consist of 8 *bits*, where a *bit* is a single binary digit.

very cleverly; but isn't the act of the programmer, the act of creating the symbolic data base, a fundamental act of abstraction and hence of intelligence? These linguistic entities tend to be like things that people would say, compact statements, but how much of a thought or worldly experience can be captured this way? So one must worry about what knowledge *is*, what knowledge is *needed*, and how to *represent* it. What is often neglected is the question of how that knowledge can *come to be known*. One way is to hire a team of knowledge engineers who spend months or years codifying the knowledge of an isolated domain into a formal system accessible to the machine. In the long run for AI, this has to be a losing proposition: how long would it take *you* to write down everything you know? And that only counts the things that you know you know and you know how to express. There's no substitute for experience.

Experience is the only possible way to amass a data base that can be said to have "world" knowledge. Experience must be abstracted, perhaps in many stages and many ways, to yield the data structures used by the higher processes, perhaps abstracted even to yield the very processes. The rope of mind has 2 ends: what do I need to know to be able to reason, and what can I say about what's happening; and it has to be spliced somewhere in the middle to connect the outside world with the inner one. Perception must be able to produce the data structures required for reasoning. In fact, given our meager understanding of intelligence, we can't really draw a line between perception and reason. Maybe there is none. After all, the relatively "minor" ability of perception has so far proven vexingly intractable.

Among our senses, vision is probably the richest and most important. Only vision and hearing have well-developed transducer technologies, making them readily accessible to attack by computer. The problems of hearing, particularly speech understanding, are no less than those of vision, but I happen to be more visually than aurally oriented, and vision has more obvious connection to geometry and topology, so it was vision that I found myself working in. Also, there was a vision group at Stanford, but no speech effort

Introduction

This work is mainly about 2 topics in computer vision:

- Edge detection
- Applications of geometrical methods.

The "geometrical methods" are those of modern differential topology and geometry.

I came to do this research in pursuit of the eventual goal of understanding and building intelligence.

An intelligent creature, whether of flesh or metal, must be able to know what is going on around it, and do something about it. Those are the peripheral functions: perception and action. These are certainly necessary, but aren't they rather minor in comparison to the "higher" functions involved in thinking, feeling, learning, language, etc.? This is an interesting question, but it isn't just this simple necessity of perception that led me to its study.

A great deal of artificial intelligence (AI) research studies the higher functions, and with varying degrees of success, tries to duplicate them. I find a curious thread running through much of this work: the manipulation of linguistic entities. People have long said that the main thrust of AI is symbol manipulation, and indeed it seems that to be smart you should be able to transform data into abstractions, and abstractions are symbols, which in some sense are linguistic entities, abstractly at least. The linguistic entities of AI tend to be statements with a great deal of meaning to the programmer, such as (DUCK IS-A BIRD), but the machine hasn't the least interest in what the symbols stand for in the w rld. The AI program endows the machine with a means to manipulate these symbols, perhaps

ought to be called an edge. This corresponds to a boundary between regions of uniform intensity measurement, uniform at least near the boundary. Very little attention has been paid to any other definition of "edge," despite the fact that close observation of images reveals that step edges between uniform (strip) regions are exceedingly rare. This is not to say that edge detectors built to detect step edges don't find real edges; indeed they often do, and indeed they often make grievous errors.

The term "edge" has been fairly widely abused, and we will continue that tradition here. One type of edge is that resulting from the boundary of some object. There are also edges which are merely boundaries between surface features. There are *local* edges and *global* edges, which are frequently called *contours*. Local and global are relative terms, and we mean them in comparison either to image or grid size. A local part of a curve, for example, would be well approximated by a straight segment in the given grid size. Thus another way of looking at the difference between local and global is related to manageable and unmanageable search problems, since locally all possible curves can be represented as all possible line segments on a coarse grid, while globally the space of all possible curves is vast.

Local edge detection

We will not attempt to give a mathematically precise as well as operationally general definition of "edge" here. Properly, to do so one would study the imaging process as well as real images. [Herskovits and Binford 1970] did so to a limited extent, presenting essentially 1-dimensional results. Essentially, what people have been looking for as edges are places with a large gradient, or places which resemble a step function in cross-section. So-called "roof" edges, modelled as a discontinuity in 1st derivative have been sought as well. It turns out that a number of different outlooks on how to look for these features lead to essentially the same computational technique, viz. convolution with some kernel followed by thresholding. (Strictly speaking, it is usually cross-correlation which

is implemented, but since the families of kernels involved are complete under inversions, we take liberties with the term "convolution.")

Spatial differentiation and gradient estimation

If edges are places where things change fast, then the obvious way to look for them is by performing a spatial differentiation. This may be done by some discrete analog of the gradient, which is implemented by convolving with a kernel of small support. The smallest possible support for a differentiation is 2 pixels, and in such a case the convolution is often thought of as taking adjacent pixel differences, or first differences. Larger supports allow more creativity in the choice of the convolution kernel defining the differentiation, and provide the benefit of improved noise behavior. A great many authors estimate gradient or "steepness" by computing adjacent pixel differences. [Martelli 1972, Martelli 1973] and [Turner 1974] are examples of the latter. Another way to think of the gradient is as a derived parameter of fitting a plane to the data. For sufficiently symmetric supports, this can also be implemented as convolutions. In fact many outwardly sophisticated techniques have as their core the estimation of gradient.

Template matching and matched filtering

A popular way to look for features is with a matched filter or template, and this is quite common for step edges. Again the cross-correlation with the template, or the space domain realization of the filter are implemented as convolutions. The idea is that the "template" (the convolution kernel) is an ideal case of the feature one is seeking, and one looks for large values of the correlation as indicating the presence of the feature. The term "template-matching" often suggests that the vector space projection analysis of the process is at best a secondary consideration. Examples are the operators of Sobel [Duda and Hart 1973] and [Kirsch 1971], as well as many others (further examples can be found in [Abdou 1978] and [Rosenfeld and Kak 1976]). The matched filter approach is operationally the same, but includes the analytical idea that as a consequence of the

Cauchy-Schwarz inequality, the maximum response for normalized data occurs when the data is the (complex conjugate of the) template. [Duda and Hart 1973] provide a more detailed discussion of the ideas of spatial differentiation, gradient estimation, and template matching, with a slightly different viewpoint. [Shanmugam, Dickey, Green 1979] seek a slight generalization of the matched filter, in the sense that the filter must be strictly bandlimited and the objective is to maximize the power of the step response in a given space interval.

Locally, i.e. at a single point of the convolution result, the integration against the kernel can be thought of as orthogonal projection onto a 1-dimensional subspace of $\mathbf{R}^n$, where n is the number of pixels in the support of the kernel, and the projection is with respect to the usual inner product on $\mathbf{R}^n$. If there is more than 1 subspace involved, i.e. more than one convolution, then one has components which can be thought of as components of a vector in the space spanned by the subspaces. Then one can compute a magnitude for that vector (so as to get a number representing "edgeness" for thresholding). The magnitude may be in the Euclidean norm

$$\|x\| = (\sum v_i^2)^{1/2},$$

or in some other norm, such as the max norm

$$\|x\| = max\{v_i\},$$

or the sum norm

$$\|x\| = \sum |v_i|.$$

Best edge fit and optimal estimation

The simplest edge model, a translated step function, has 3 parameters (for a 2 dimensional picture). These might be, e.g., angle, left height, and right height. With enough

normalization, these can be reduced to the single parameter of angle. Template matching methods use a separate template for each angle considered. But one can also try to determine the angle that best accounts for the data. Furthermore, the model may have more parameters, and there may be statistical information available.

The simplest type of best fit problem occurs when the model space is a linear subspace of the data space, which is an inner product space. In that case, the best fit is obtained by orthogonal projection to the model space. This is a very common method for fitting functions in 1 dimension, based on the observation that translation in space is equivalent to multiplication by a complex-valued function of frequency in the frequency domain, so that all the translates of a given frequency component make up a linear subspace. In 2 dimensions, though, matters are complicated by the presence of rotations, so that while the same artifice applies to translations, the Fourier equivalent of rotation is still rotation, and the set of all rotations of a component is no longer a *(1-dimensional) linear subspace*, so direct orthogonal projection is no longer applicable. Hence many methods which seem very clever for 1 dimension fail for 2 dimensions. However, this nice property of Fourier transforms for 1 dimension can be thought of as a special case of a more general principle, which may be of use in inventing best fit methods. Specifically, one way to restate the spectral theorem [Halmos 1957, Halmos 1963] is that any normal operator in a Hilbert space is unitarily equivalent to a multiplication. For our purposes the Hilbert space can be taken to be $L^2(\mathbf{R}^2)$. Then the spectral theorem can be interpreted to say that given a normal operator A, we can find some isometry $U : L^2 \to L^2$ and some function $\varphi \in L^2$ such that $U^{-1}AU(f) = \varphi f$ for all f simultaneously. If A is a translation operator, the Fourier transform is such a unitary transformation, as we mentioned above. According to the theorem, there is some isometry of $L^2(\mathbf{R}^2)$ which will transform rotation into complex multiplication. Using that isometry like a Fourier transform, one could use projection methods to find best fits. Even better would be a transform that worked for translations and rotations at the same time, but that is impossible because translations do not in

general commute with rotations (as would clearly be necessary for the existence of such a transform because multiplication is commutative).

A slightly more general best fit problem occurs when the model set consists of an n-parameter family of functions, and the object is to find a member of the family which minimizes some error measure with respect to the datum. If the family is differentiable, it can be thought of as a submanifold of the ambient space. Frequently the error measure is a metric on the space, and then the problem is seen as one of finding the closest point of the model manifold to the datum. In the case of estimation, there is a probability distribution involved, and one seeks a set of parameters minimizing the *expected* error.

[Altes 1975], [Hueckel 1971,Hueckel 1969], [O'Gorman 1976], [Abdou 1978] find best fit edges. Altes uses essentially the 1-dimensional Fourier method described above. Hueckel and O'Gorman minimize the distance between the projection of data and parametrized model onto a truncated orthonormal basis, deriving the "optimal" parameters. However, both the number of parameters and the number of terms in the series are too small to allow good performance. Altes uses a more realistic edge model (in 1 dimension), but his results are not readily generalized to 2 dimensions. Abdou finds the best fit edge by what is essentially an exhaustive search over a slightly more general but still too simple model space, namely linear ramps between constants.

When the parameters one is seeking are the coefficients in an orthonormal basis, the parameters can be obtained simply by taking the inner product with the basis elements.

Higher order derivatives

Methods that rely on estimates of the gradient, or whose response is largely determined by the gradient cannot distinguish smooth transitions from abrupt ones. In the Hueckel and O'Gorman approaches, for example the early term(s) in the expansions are essentially the gradient. One approach to this problem is to use a preprocessing step which takes

linear functions to 0. This idea is advanced by [Binford 1981] in the form of "lateral inhibition," and in fact operators modelled on second and higher order derivatives will have this property. (It's interesting to consider just how many such operators there are. Suppose the operator support is n pixels. There are then n linearly independent such operators. Requiring that all operators take constants to 0 is a linear constraint, and that they take all linear functions to 0 is 2 more linear constraints, so there are $n-3$ linearly independent operators fulfilling the constraints. One may impose further constraints by requiring various symmetries, and each discrete symmetry will reduce the dimension of the operator space by 1. For large supports, it is clear that there are many candidate operators.) The second derivative in the calculus of several variables is the Hessian, which is a matrix. Its algebraic invariants are the geometric invariants of the original function viewed as a surface. Various combinations of its components (taken linearly and nonlinearly) can be used as 2nd derivative operators. If an edge is sought at suitably defined maxima of the gradient, then for a 2nd derivative operator, one seeks zero-crossings. [Marr and Hildreth 1979] use an approximation to the Laplacian, which is the trace of the Hessian. [Dreschler and Nagel 1981a, Dreschler and Nagel 1981b] use the determinant of the Hessian. [Beaudet 1978] computes rotationally invariant derivatives up to 4th order. [Canny 1983] takes an optimal estimation approach to the zero-crossing of 2nd derivative problem of [Marr and Hildreth 1979], using criteria of detectability and localization in a variational formulation.

Approximation and representation of image function

One of the drawbacks of the methods we have been describing is that a very few parameters are derived by some kind of local projection. The parameters are chosen for semantic interest, but while they respond well to intended features, the same is often true for unintended features. We have the following situation. Let X be the space of all local images, and $F \subset X$ the features one is seeking. Perhaps this is done by some map $\varphi : X \to \mathbf{R}$. One designs this map so that $\varphi(F) \geq \Theta$, for some threshold Θ, and one

would like to be able to infer that if $\varphi(x) \geq \Theta$, then $x \in F$. Clearly, to do this, one must have some information about $\varphi^{-1}(-\infty, \Theta)$, but this is surprisingly often neglected.

Another way to think of this is that the few semantically derived parameters actually do not provide enough information to understand the structure of the image intensity function, even locally. Now, of course the pixel values constitute complete information, but it is not directly usable. One approach, then, is to seek a local representation for the image data which is appropriate for the questions one wishes to resolve with further processing. Approximating the Hessian is such a process, since that can be regarded as finding the best local quadratic approximation, just as computing a gradient can be viewed as finding a planar approximation. [Prewitt 1970] computed her gradient parameters based on a planar fit. In the same vein, [Haralick 1980] fits planes to the data and defines edges as boundaries between maximal domains of fit, relative to an error measure. Planar fitting is very crude so he [Haralick 1981] proposes polynomial fitting as an extension. [Beaudet 1978] is motivated by fitting a truncated Taylor series, though the semantics he ascribes to his operators are somewhat naïve. [Hsu, Mundy, Beaudet 1978] use a quadratic fit, based on Beaudet's techniques. [Altes 1975] is put forward as essentially a spline fit.

Global edge detection

The Hough transform [Hough 1962], [Duda and Hart 1971, Duda and Hart 1972, Duda and Hart 1973] is a technique to find collinear sets of feature points over an entire image. This can be applied in complete globality, i.e. over the entire image at once. [Ballard and Sklansky 1976] [Shapiro 1974, Shapiro 1975, Shapiro 1978], and others use generalizations of the method to look for other 1 dimensional objects.

Frequently, the term *linking* is used synonymously with global edge detection. Linking consists of making lists of local edge elements connected head to toe, each list corresponding to an extended (global) edge. This is the most common global edge detection method,

dating back at least to [Roberts 1963], and including many others (e.g. [Horn 1972], [Binford 1970], [Nevatia and Babu 1978]). These methods differ primarily in the predicates used to determine whether to join a particular edgel into a contour. A major difficulty stems from the fact that the linking proceeds only after irreversible decisions are made about local edges, e.g. limiting each pixel to having an edgel of unique orientation, or making a binary decision about the presence of a local edge. The type of information available to the linking, which generally proceeds locally, is inadequate for many situations.

An improvement on the linking method is advanced by [Montanari 1970, Montanari 1971] and [Martelli 1972, Martelli 1973]. Here the prior local commitment is less extreme, and dynamic programming or heuristic graph search methods are used to find optimal paths with respect to some figure of merit. The figure of merit, a global parameter, replaces the local predicate as the contour selection method, and likewise as the main artistic element.

The "relaxation" methods propounded by [Zucker, Hummel, Rosenfeld 1977] and [Rosenfeld, Hummel, Zucker 1975] attempt to find the contours globally, in parallel, and without excessive initial commitment. The process depends on a local pairwise reinforcement-inhibition process between edgels. The art is in choosing the reinforcement process. Explicit global edges are not produced, but presumably the process terminates with sets of edge points which are both connected and of a desired minimum length, which are then readily identified.

Region growing

We motivated edge detection as a means to region finding. Why not just find the regions directly? Many people have tried doing just that. The advantage is that one is dealing with a global object, so the problem of linking is (or seems to be) avoided. Rather than deciding whether an edge separates 2 points, one must decide whether 2 points belong to the same region. Seen thus, the difference is mainly one of (linguistic) semantics. The

data structures reflect regions, not edges, as do the algorithms. Consequently, despite the conceptual equivalence with edge finding, different approaches, harder to express in the edge detection paradigm, are developed. The simplest method is based on segmentation simply by intensity or color value. [Brice and Fennema 1970, Fennema and Brice 1970] take this as their starting point, and then try heuristics to clean up. [Ohlander 1975] segments based on dividing bimodal histograms of several color parameters. [Shafer 1980] builds on Ohlander's work. [Somerville and Mundy 1976] use a technique based on more sophisticated reasoning. They grow regions based on the uniformity of an approximation to the normal to the image intensity function. [Kirsch 1971] defines regions based on thresholding a "contrast" (gradient) function.

In the following, I have attempted to provide a critical guide to the literature in segmentation. The list of works reported on is by no means exhaustive, but it is intended to include the most influential works as well as some others representative of the field. In addition to summarizing each work, I have usually tried to put it into some perspective, which is to say that I have included many of my own reflections. I hope that the boundaries between the two are discernible enough.

General Works

Prewitt 1970

"Object Enhancement and Extraction"

The paper is concerned with the entire image understanding problem:

- image formation
- image restoration
- enhancement (including edge enhancement)
- object extraction

The author provides a fairly extensive bibliography (237 references) of literature at that time (much of which is still germane).

The work is fairly sophisticated mathematically. E.g., Prewitt considers the Laplace, Mellin, Fourier, and Hankel transforms, moments, Haar-Walsh functions (cf. [O'Gorman 1976]), Chebyshev polynomials, point spread function (PSF), line spread function (LSF), edge spread function (ESF), modulation transfer function (MTF), and phase transfer function (PTF). She also discusses resolving power and restoration, including "super-resolution" for restoring images which have been degraded by a convolution (referencing e.g. [Slepian and Pollak 1961, Landau and Pollak 1961, Landau and Pollak 1962] and applications).

Edge enhancement

A section devoted to edge enhancement discusses the gradient, generalized derivative, Laplacian, and discrete approximations to gradient.

As one means of obtaining an estimate of the gradient, she introduces the 3×3 (now so-called) "Prewitt operator:"

$$\begin{matrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{matrix} \qquad \begin{matrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{matrix}$$

This is used in a method of estimating the gradient by fitting a quadratic surface to data on a 3x3 square. The masks give $\partial/\partial x, \partial/\partial y$ for that surface directly from the data. This is exactly the method used by [Haralick 1980] for facets. Similarly, one can use the same idea for a 4×4 fit or a Laplacian.

She also discusses oriented edge masks, e.g.

$$\begin{matrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ -1 & -1 & -1 \end{matrix}$$

as approximations to the gradient ("compass gradient"), and gives some examples of their use.

A discussion of modified "crispening" (Laplacian) operators is presented, as well as of line enhancers (which are basically templates, i.e. matched filters).

Frequency filtering

Low, high, and band pass filtering is considered.

She discusses templates, matched filtering, and cross correlation for feature detection.

A good discussion of thresholding is presented.

The paper is an excellent overall survey of the then-existing methods for feature extraction, and in particular edge detection. By and large, the intervening years have seen only minor improvements, so the analysis she presents is still relevant today.

Evaluation

Aside from frequency domain filtering, the methods presented, including the "Prewitt" operator, are completely local with small support—in her words, "context-insensitive." Consequently, global structures cannot contribute to the edge finding process, and the derived image description is limited to 1 or 2 local parameters which provide inadequate description of the image intensity function for all but especially regular images.

Unlike most gradient estimation or template matching operators, the Prewitt operator is based on a well-defined process--the best fit of a plane. The gradient by itself is not sufficient for edge detection, since no discrimination is made between smooth and abrupt transitions, although plane fits can be used in more sophisticated ways (see e.g. [Haralick 1980]).

Davis 1973

"A Survey of Edge Detection Techniques"

The author presents some discussions of prior edge detection techniques:

Parallel edge detection

Herskovits and Binford 1970

linear *vs.* nonlinear operators (nonlinear: mainly Rosenfeld, Hummel, Zucker 1975)

texture edges

Griffith 1970, Griffith 1973a, Griffith 1973b

Hueckel 1971, Hueckel 1973

Chow and Kaneko 1972

Sequential edge detection

Martelli 1972 Montanari 1970

"Guided" (top-down) edge detection

Kelly 1971

Harlow and Eisenbeis 1973

Shirai 1973

He discusses and criticizes what was done very tersely. There are no particularly deep or sophisticated analyses; nevertheless this work provides a useful first tour or refresher of some of the more significant work in the field. One can detect a subtle and not surprising bias toward Rosenfeldism.

Kanade 1978

"Region Segmentation: Signal vs. Semantics"

A survey of image segmentation is presented, based on the paradigm: Image → Picture Domain Cues → Scene Domain Cues → Model → Instantiated Model → View Sketch → Image ..., which may be iterated. A distinction is made among the categories of signal, physical, and semantic knowledge.

A large number of works are briefly surveyed, and categorized according to how they fit into the above paradigm. For example, many methods use only signal level knowledge, and hence, in this paradigm, can provide at most a segmentation based on picture domain cues.

Evaluation

The paradigm presented can be more conventionally summarized as saying that one's goal must be to infer the 3-dimensional structure of a scene in order to model the scene and understand the image. Furthermore, one must use physical knowledge, e.g. imaging physics and geometry, to make this inference. This is hardly new or controversial. What is debatable, however, is the distinction which is made between picture and scene domain cues. The main orientation of the paper is toward region growing and splitting methods,

using fairly primitive "signal level knowledge," e.g. histograms of the image gray values. For these types of systems, the image-picture-scene-model division is clear and seems natural. But for "image understanding" in general (which the author is addressing), such an easy description does not seem justified, and no arguments are presented to persuade the reader, though in the author's defense it must be said that there were severe space limitations for a fairly broad article.

It seems reasonable that the first step in image understanding might well be to compute a description of the image data in a more useful representation, or set of representations, than is provided by the standard one, i.e. the set of pixel values. Kanade notes, in fact, that [Pavlidis 1972] defines ***segmentation*** as a process for describing the image features themselves. From this point of view, "picture cues" are features of this re-representation. (Kanade takes a more restrictive and ill-defined view; he defines "picture cues" by the examples: line segments, homogeneous regions, and intensity gradient. The last of these is properly a property of the image, but it can be argued that the first two generally cannot be extracted reliably without using knowledge about 3-dimensional structures, and that is tantamount to making inferences about the "scene domain," although admittedly historically such inferences are implicit.) But it is not so obvious that there must be a trichotomy: picture-scene-model. First, the new image representation is chosen based on physical knowledge--the knowledge that determines for what it is important to look. Whatever features are focused on in analyzing the new image representation are likely to be interpretable as features in the scene domain only in conjunction with fitting them into a model. For example, the interpretation of a narrow gradient-shaded region may depend on its connection to other regions and on some set of hypotheses about other regions in the vicinity. This might even be on the level of deciding whether the region is an object limb, a surface, a highlight, or even whether it should be regarded as a separate region at all. One can readily envision a Waltz or Zucker type relaxation process occurring using the semantic relations of a model to interpret part of an image representation as a scene

ich expansions cannot avoid truncation error. However, if the sampling kernel is taken (say) C^∞, its linear combinations (i.e. the functions $\sum \alpha_i \psi_i$, where the ψ_i are discrete anslations of the sampling kernel ψ) become prime candidates for an expansion series. hese can be frequency (or sequency) ordered. Expansion in terms of such functions has een extensively studied; use of truncated series fitting is worth investigating.

urner 1974

Computer Perception of Curved Objects Using a Television Camera"

Ve are concerned here only with the edge finding aspects of this work

he author gives a brief critical synopsis of earlier line finding work:

Binford-Horn [Horn 1972]

Griffith 1970

Herskovits and Binford 1970

Hough 1962

Hueckel 1971, Hueckel 1973

Kelly 1971

Murphy 1969

O'Gorman and Clowes 1976

Pingle 1966

Pingle and Tenenbaum 1971

Roberts 1963

Shirai 1973

Tenenbaum 1970

he edge finder Turner employs is very simple, using the first difference of adjacent pixels, ollowed by thinning, and further by a local tracker (inchworm).

short review of curve segmentation is provided.

assumes that the digitization process takes place by averaging over a square pixel sized window, i.e.

$$g_{ij} = \int P_{ij} f dA,$$

where

f =image irradiance
P_{ij} =unit 2-dimensional pulse at the point (i, j)
g_{ij} =the sampled output,

then the P_{ij} constitute an orthonormal set whose span is identical to the Walsh functions of order less than $I \cdot J$ (where I, J are the cardinalities of the i, j sets). The higher order Walsh functions describe exactly only what goes on within pixels, which is precisely the information lost in the digitization process, so one has a perfect match of model to data. The Walsh basis differs from the single pixel basis most notably because the support is spread over the entire region of interest, i.e. the Walsh basis has *global support.* Truncating the series therefore results in global degradation, rather than local as would be the case with the analogous action of leaving off some set of pixels.

Unfortunately, incorporating sufficient Walsh terms to utilize all the picture data is equivalent to doing a fit of a perfect edge to the sampled data with the pixel value = average intensity assumption. This becomes extremely complex as the number of pixels increases, and if lateral displacements of edges are permitted, since the discontinuous pulse convolution kernel forces independent examination of numerous cases corresponding to the edge configurations' relations to corners of pixels. O'Gorman already has to consider 2 such cases for a 4x4 operator and 6 Walsh functions. As the space grows larger, so does the complexity, so that [Abdou 1978] chooses to do an exhaustive search as his method of fit.

The advantage of everywhere differentiable functions (such as [Hueckel 1971, Hueckel 1973] uses) is that the lack of discontinuity permits a single set of equations to express the optimization problem. Of course, if one assumes a discontinuous sampling kernel,

the point spread function, one would have to preserve this property while generalizing the integration properties, which is not at all trivial. What comes to mind is using tensor products of the 1-dimensional $\delta^{(-n)}$ functions, which could be expected to have similar properties, but that would not lead to a simple description of boundaries.

The method of locating the free knots is not very clearly presented, and appears to be based on a possible misconception. In 2 dimensions the problem is of course more difficult because of the complexity of the boundary space. There is no obvious way to solve this problem.

The paper is more biology oriented than computer oriented, so understandably no consideration is given to digital processing issues, the most important of which is the effect of discrete sampling on a periodic grid.

O'Gorman 1976

"Edge Detection using Walsh Functions"

O'Gorman shows that finding edge direction by fitting a plane and then taking its gradient direction is subject to systematic error for perfect step edges centered in a square window. However, this is a consequence of the shape of the window - a circular window would not have the same problem. Nevertheless, the analysis is salient because pictures are sampled on a square grid and rectangular operators are common.

He uses the 2-dimensional Walsh functions (tensor products of square waves) as an orthonormal basis for representing the image function. In analogy to [Hueckel 1971, Hueckel 1973] he does an L^2 (least squares) fit of a perfect edge on the first 6 terms (in his Walsh expansion).

The contribution of this idea derives from the fact that the Walsh basis bears a simple relationship to the digitization process (if one assumes square pixels). In particular, if one

It is interesting to compare these with the difference of Gaussians suggested by Marr and Hildreth based on similar assumptions.

Some comparisons with human vision are made, notably between line spread functions, which are shown to be similar.

Evaluation

The work is quite provoking. The most interesting features are that it incorporates the transducer transfer function, models images as intrinsically discontinous objects in a coherent way, and uses statistical estimation for detection. Unfortunately, the generalization to 2 dimensions is not easy, and probably not as easy as Altes seems to suggest. He proposes 2 routes of "generalization." The more straightforward involves using rasters at a number of angles. Though this is not as satisfying as an intrinsically 2-dimensional approach, it may be a viable way to proceed. Significant problems that would have to be overcome include integrating all the information from the various scan lines (which could be argued to be 99% of the problem to begin with), and accounting for or using a 2-dimensional transducer transfer function. Making a true generalization to 2 dimensions poses the following difficulties. Knots are of codimension 1; i.e. they are boundaries between regions, so on a space of 1 dimension, a knot is 0-dimensional, or a point. But on a space of 2 dimensions, a knot is the boundary of a region, i.e. some curve, a 1-dimensional object. So for 1 dimension, the space of knots is 1-dimensional (since it is the space of points), but for 2 dimensions the space of boundaries is infinite dimensional (since it is a space of curves). The approach of workers in spline theory has been to generalize the intervals between knots to projections from higher dimensional simplices, leading in the 2-dimensional case to piecewise straight boundaries, but this seems to be inadequate for a natural description of the boundary. The 2-dimensional analog of the delta function at a knot is a delta function whose support is a boundary. Since the main virtue of using the $\delta^{(-n)}$ expansion is the simplicity of convolution with

where $u^{(-n)}$ is defined analogously to $\delta^{(-n)}$, since $u * \delta^{(-n)} = u^{(-n)}$.

Working in the Fourier domain, he introduces a derived set of normalized basis functions, and shows how to estimate the coefficients in the expansion in the case of a single knot of known position. For multiple free knots, he proposes using techniques from detection and estimation theory, based on a statistical model of knot location. However, the approach is predicated on the use of a matched filter to locate the knots, which appears to be doomed to failure because the basis is not orthogonal.

The core of the author's method uses filters to estimate coefficients or detect complex patterns. Based on filter complexity considerations, he argues that these filters should all have approximately equal space-bandwidth products. These arguments are related to implementation issues, and for the digital case would be related to cost. One must keep in mind, however, that a major consideration of the work is a theory of human vision. In order to achieve a set of filters with the desired property, he seeks a set of transducer transfer functions to incorporate into the imaging transfer function U. Although it is not stated in the paper, one can think of this as a convolution preprocessor which allows further processing to be done by filters all having the same space-bandwidth product. He uses one particular way of obtaining a constant space-bandwidth product, viz. $V_n(\omega) = \alpha_n V_{n-1}(k\omega)$ for all n with a fixed constant $k > 1$, where α_n is an arbitrary proportionality constant and

$$V_n(\omega) = \frac{U(\omega)/(i\omega)^n}{\|U(\omega)/(i\omega)^n\|},$$

where $\|\cdot\|$ signifies the L^2 norm. Although this is a simple way to get a constant space-bandwidth product, it is not the only way: e.g., a different k could be used for each n. In any case, using this assumption, he arrives at a set of log-normal transducer transfer functions, i.e. functions of the form

$$U(\omega) = A\omega^{\nu} e^{-\rho(\log\omega)^2}.$$

the color space. From this point of view, 3) is not really possible (unless one wants to be ad hoc), 2) is unsophisticated (though it may be adequate in many cases, but won't maximize S/N). It might be computationally efficient to choose not a basis, but a larger spanning set. 1) therefore is the way to go. Note that there may be more than one metric which is worth using simultaneously. It is also worth investigating the differences between using a metric such as

$$d(p,q) = \|p-q\| = \sqrt{\sum (p_i - q_i)^2}$$

and using a function

$$\rho(p,q) = \big|\,\|p\| - \|q\|\,\big| = \left|\sqrt{\sum p_i^2} - \sqrt{\sum q_i^2}\right|.$$

Notice the latter is like the intensity difference.

Altes 1975

"Spline-like Image Analysis with a Complexity Constraint. Similarities to Human Vision"

The author proposes modelling a (1-dimensional) picture as a finite sum of basis functions which are integrals of delta functions:

$$f(x) = \sum_{n=0}^{N} \sum_{m=0}^{M} f_{nm} \delta^{(-n)}(x - x_m),$$

where $\delta^{(-n)}$ is the nth integral of the unit Dirac delta function, $0 \leq n < \infty$, and the x_m are free knots. Splines can be viewed as such sums with $1 \leq n < \infty$ and smoothness conditions imposed at the knots, hence the paper's title. Including the point spread function, u, of the imaging system yields

$$f(x) = \sum_{n=0}^{N} \sum_{m=0}^{M} f_{nm} u^{(-n)}(x - x_m),$$

Nevatia 1977

"A Color Edge Detector and Its Use in Scene Segmentation"

Nevatia's goal in this work is to define a Hueckel operator for the 3-color domain.

A review of color space representations is presented.

He states there are 3 ways to look for color edges:

1) Choose a metric in the color space and look for discontinuities

2) Choose a basis and look for edges in the projection to each basis element separately

3) Do 2) but require uniformity to use 3 components together

He chooses to do 3).

However, what he actually proposes doing is minimizing the sum of the squares of the errors of the individual color component Hueckel fits. This is exactly equivalent to choosing an inner product on the color space such that the 3 color components are all orthogonal, then using the metric induced by the inner product, i.e. the Euclidean metric. This, as he points out, is equivalent to minimizing the individual components separately. Doing so, though, would lead to 3 fits for the 3 components which might have nothing whatever to do with each other (since one is not looking for the single edge that leads to all the data, but independent edges for 3 sets of data. Therefore, he imposes the additional constraint that the inclination angles for all 3 solutions must be the same, i.e. he adds the 2 equations $\alpha_1 = \alpha_2 = \alpha_3$. However, computing this angle is not easy, so instead he takes a weighted average of the 3 independent solutions (i.e., without the single angle constraint).

The idea of "best" fit implies a metric, since one must have a way to measure how good the fit is. Hence there is no way to avoid (explicitly or implicitly) choosing a metric for

Evaluation

These papers contain some mathematical inaccuracies, which in themselves are not very important, but whose presence brings into question other mathematical claims which are not proven. An example of an inaccuracy is the statement that "The set of all continuous functions over [the closed unit disk] is a Hilbert space." Since a Hilbert space is defined to be a complete normed inner product space, the statement is false because the space in question is complete in the sup norm, where there is no inner product, but not complete in the inner product space L^2, which is the one Hueckel is using. One might then be more skeptical of the claim that the basis functions he settles on are the unique solutions of some unspecified set of "functional equations."

The main contribution here is to approach the best edge fit problem in a tractable subspace, thereby transforming an essentially combinatorial problem into an analytic one. The particular implementation of that idea, however, suffers numerous shortcomings.

Several criticisms have appeared in the literature. [Abdou 1978] argues that the truncation of the orthogonal series introduces excessive error, especially for thin lines, and that unjustified assumptions are made in the optimization procedure. [Shaw 1977, Shaw 1979] makes a similar criticism of the optimization. [Davis 1973] complains that no attempt is made to relate performance to the image noise process.

Experience using the operator shows that regions of smooth shading result in multiple firings, while regions busier than the size of the operator have missed edges and poor parameter values. These failures are a consequence of using a poor model for the underlying image intensity function. The edge and edge-line models are unrealistic, especially for the support area of the operator. The difficulty can be traced to the fact that in the spaces considered, ideal edges and linear functions are not mutually orthogonal.

Unfortunately, no analysis exists, either here or elsewhere, of the error one incurs by using such simplistic models.

to minimize $d(\pi_k f, \pi_k E^*_{\theta,p})$ with respect to θ, p. Since the basis is orthonormal, this can be done componentwise. This is computationally efficient because the series is truncated at a point which allows a closed form solution for the least squares problem. The line paper uses essentially the same method, with additional parameters to allow for an ideal step edge-line, i.e. a sum of 2 parallel ideal step edges. The method can equivalently be thought of as fitting the best function from the fixed subspace S_k to the data and finding the best edge fit to the function. (This is a consequence of orthogonalities of various subspaces).

The orthonormal expansion used consists of polynomials in x, y with a uniform radial weighting function $\sqrt{1-x^2-y^2}$. For the edge (old) operator, 8 polynomials up to degree 3 are used, while the edge-line (new) uses 9 polynomials up to degree 4 (neither set spans the space of all polynomials up to their maximum degree). What, if any, classical set of orthogonal polynomials these correspond to is not stated and not immediately evident, since the definition of the basis functions is presented in a complex way. The orthogonal functions are related to a Fourier-Bessel basis, since $x = r\cos\theta, \quad y = r\sin\theta$, and the r polynomials can be thought of as approximations to the Bessel functions one obtains for a radial Fourier transform. It is not stated how the basis functions were derived, however.

The edge/no-edge decision is based on the "angle" between the projections of the data and the best fit edge in the truncated space S_k. I.e., he thresholds on the value of

$$\frac{(\pi_k f, \pi_k E_{\theta,p})}{|\pi_k f| \cdot |\pi_k E_{\theta,p}|}.$$

This suffers from the common problem that little analysis is devoted to the possible picture functions $\pi_k^{-1}(\pi_k E_{\theta,p})$, which are going to look like edges to this operator. In particular, the average gradient plays a large role, and the decision criterion therefore tends to respond to areas with large average gradients over the support.

Evaluation of line finding

This was an early effort. It probably is not bad for straight lines, though it seems to miss a lot. Curved edges or complex scenes are not handled, and many *ad hoc* methods are used.

The technique presented here has no hope of working where there are wide variations in smooth shading gradients, since the thresholds are global, and the gradient operator cannot discern whether the signal is from a smooth gradient or a local step.

Of course, it must be stressed that Roberts broke ground in the use of his gradient operator, as well as in the use of homogeneous coordinates, the fitting of 2-dimensional data to 3-dimensional models, and in line following.

Hueckel 1969, Hueckel 1971, Hueckel 1973

"A Local Visual Operator Which Recognizes Edges and Lines"

[Abdou 1978] presents a detailed analysis, to which we direct the reader rather than repeat the same points.

The method involves finding the parameters of the best fitting ideal step edge in a disc-like region of 32 to 137 pixels. The fitting is done in the spirit of the Rayleigh-Ritz method of finding approximate solutions to variational problems (see, e.g. [Morse and Feshbach 1953]). Using a fixed orthonormal basis for the function space of interest, and a fixed truncation of the orthonormal basis, he finds parameters to minimize the L^2 distance between the projections of data and ideal edge in the finite dimensional space spanned by the truncated series. I.e., let ψ_i, $i = 1, \ldots, \infty$ be an orthonormal basis for L^2. Let $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ be the picture (data). Let $E_{\theta,p}$ be an ideal step edge of orientation θ centered at $p \in \mathbf{R}^2$. Consider the space S_k spanned by the first k basis vectors, $\psi_1, \ldots, \psi_k$, and let π_k be the orthogonal projection onto that space. Then the idea is

- correlate (i.e. sum) along lines of length = 5, for values of $\theta = n \cdot 45°$.
- threshold on the ratio $\frac{best}{worst}$ of the line values, yielding edges.

Linking

- connect edgels if:
  1) they lie in contiguous 4x4 squares.
  2) they are related by a $< 23°$ change in direction.
- eliminate singletons.
- apply an ad hoc cleaning processes for small triangles, quadrilaterals, and spurs.

Curve representation and segmentation

- least squares fit straight lines to linked sets.
- uses sequential (updating) method of fit.
- first done on connected edgels.
- choose a random starting point, then proceed until:
  1) a branch is reached, or
  2) an error threshold is exceeded for the line fit, in which case back up until the local angle to the fit is cut by 1/2.

The remainder of the paper is concerned with the recognition and display of polygonal 3-dimensional objects.

Local edge detector

He first takes the square roots of the pixel values, on the basis of psychophysical evidence which he cites. In a 2x2 pixel window, let the square root values be

$$\begin{matrix} a & b \\ c & d \end{matrix}$$

The edge measure is then defined by

$$\varphi = \sqrt{(a-d)^2 + (b-c)^2}$$

This is proportional to the gradient magnitude of a least squares fit plane (e.g. [Haralick 1980]). I.e., if F is the best fit plane,

$$|\nabla F| = \frac{1}{\sqrt{2}}\sqrt{(a-d)^2 + (b-c)^2}$$

Roberts cautions that his line finder "makes mistakes in complex pictures and is a complex special-purpose program demonstrating very few general concepts." One must keep in mind that this was a pioneering work and his main interest was higher level model matching.

We summarize the operations performed in line finding in the following lists.

Edge detection process

- $\Phi = R_\nabla(P)$ (do "Roberts cross" operation, i.e. compute $|grad|$).
- take *max* on each 1×1 square of a tesselation.
- threshold.

Local Methods

Best fit techniques

Roberts 1963

"Machine Perception of Three-Dimensional Solids"

This is a seminal work, often cited as the first serious attempt at a functioning computer vision system.

The research described seeks to match pictures of a narrow class of prismatic solids to stored models, starting from raw picture data. There is a wide range of issues which the author had to address to achieve this; since we are concerned here with segmentation, we ignore most of the other contributions of the paper.

The central task the program performs is to match a wire frame model to derived wire frame data. An important part of this consists of vertex matching. To this end, he tries to fit n-point data (2-dimensional) to an n-point model (3-dimensional) by finding the best transforms H, D in homogeneous coordinates such that

$$AH = DB + \epsilon,$$

where

A = n points (x, y, z, w) from the model

B = n points (y, z, w) from the data (uses x as projection axis)

H = 3×4 homogeneous perspective transform

D = Diagonal $n \times n$ scale matrix

ϵ = error matrix

He solves this as a least squares problem.

feature. In the shape-from-shading paradigm, for example, one is hard put to identify any stage as "picture domain cues."

In summary, the paradigm presented is a useful one for discussing extant image understanding systems, and is particularly clear for those based on rudimentary image characteristics. One must be careful, though, not to be misled into a dogmatic adherence to the paradigm presented, since it seems likely, perhaps necessary, that it is inadequate as a description of the type of system required to do successful image understanding in unrestricted environments. The survey is readily accessible as well as concise; it is recommended as a good entry into a fair portion of the segmentation literature.

He discusses the (φ, s) representation of plane curves, defined by

$$\varphi = \text{tangent direction, and}$$
$$s = \text{arc length.}$$

Then

$d\varphi/ds = \kappa$ is the curvature, and
$d\varphi/ds = constant \Leftrightarrow$ the curve of φ vs. s is a straight line
$\Leftrightarrow \varphi =$ a linear function of s.

Curves are then found by fitting straight line segments to the (φ, s) data.

Abdou 1978

"Quantitative Methods of Edge Detection"

This work is concerned solely with local operators.

The author presents a review of several such operators:

Roberts 1963

Sobel [Duda and Hart 1973]

Prewitt 1970

Compass gradient [Prewitt 1970]

Kirsch 1971

3-level, 5-level [Robinson 1977]

Hueckel 1971, Hueckel 1973

It is interesting to note, perhaps as a comment on the literature in general, that Abdou presents 8 different 3×3 convolution operators. With a support of 9 pixels, there can be only 9 linearly independent 3×3 operators (since they make a 9 dimensional vector space). The 8 presented are in fact linearly independent, and the further inclusion of an

operator which picks out a single pixel value (*the trivial operator*), e.g.

$$\begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{matrix}$$

would result in a set which spans the entire space.

He evaluates the performance of convolution operators on perfect step edge visual input, assuming square pixels and area-proportional sampling (i.e. the pixel value $g(p)$ is defined by

$$g(p) = \int_{R(p)} f dA,$$

where $R(p)$ is the (square) pixel support region). This leads to complicated formulae for pixel values from rotated edges.

He discusses statistical aspects of edge detection and evaluates the 2×2 and 3×3 operators with respect to statistical performance. E.g., probabilities of detection *vs.* false detection for various S/N are compiled.

A discussion of edge detection as pattern classification is also presented, including the application of the Ho-Kashyap algorithm to the problem.

A review of statistical methods is presented, focussing on various methods of hypothesis testing:

Bayes decision rule

Neyman-Pearson criterion

minimax criterion

All evaluations are based on assuming the input to consist of a perfect step edge plus simple (usually Gaussian) noise. Unfortunately, real data rarely have perfect step edges and usually have non-constant areas which are not edges. (See e.g. the review of [Canny 1983] for a more detailed discussion of this assumption.)

An analysis is presented of the effects of Gaussian noise for linear masks.

Abdou uses Pratt's figure of merit to test the various convolution operators. The best performers are the 3-level and Prewitt operators (which are essentially the same). (Pratt's figure of merit is defined as follows. The input is a perfect vertical ramp edge, i.e. a function only of x, having a cross-section of constant-ramp-constant, i.e. a constant part connected by a linear part to another constant part. The variable parameters of the input are the contrast (the difference between the 2 constant values), the slope (of the linear transition ramp), and the standard deviation of additive Gaussian noise. The figure of merit is then defined by a formula based on parameters of the output error. Also, an analogous version is presented for edges at a 45° angle to vertical.)

For convolutions with square support, he analyzes the effects of mask size, center-weighted masks, and local adaptive thresholds.

Abdou proposes 2 new edge operators: 1- and 2-dimensional ramp best fits, resp. The idea for the 1-dimensional case is to fit an ideal 1-dimensional ramp edge to the data for all possible ramp sizes (with discrete end points). Results for each size are given in closed form, but the various sizes must be considered separately to determine the best among them. The 2-dimensional ramp best fit proceeds in the same way as the 1-dimensional, but he also considers all possible orientations. These he limits to multiples of 15°.

There are several appendices:

Analysis of the Hueckel operator (fairly good)

Orthogonal transformation in edge detection

(the beginnings of a DFT method of edge detection)

The Herskovits algorithm (not a very enlightening discussion)

Derivations of Eqs. 3.29, 3.31, 3.32 (some statistics)

Experimental results (pictures)--not very informative, extensive or useful.

He only provides binary edge maps of 3 pictures. One can't really

see what is happening locally (the pixels are too small to be seen).

Evaluation

To the extent that the local edge ramp hypothesis is valid, the ramp fitting method may work, though it is essentially equivalent to applying various slope masks in various directions. This is a rather inelegant approach since a best fit must be performed for each possible ramp width and angular orientation, with the optimum found by exhaustively comparing the error parameters for all the fits. One advantage over gradient operators and other best fit operators is that the present method can be used to reject regions of smooth shading if the all-ramp condition is rejected as not an edge. The main virtues, then, stem from the enlargement of the space of possible features to include the ramp edges. However, the Haralick "facet" model is more general, no more expensive, more elegant, and probably more effective, though probably also inadequate (see review of [Haralick 1980]).

Beaudet 1978

"Rotationally Invariant Image Operators"

The author is interested in finding a least square polynomial approximation to image data. The coefficients of the monomial terms are computed via convolution.

The starting point is to consider the polynomial to be fitted as a truncated Taylor series. The coefficients are found as in a normal least squares problem, but are taken to represent the derivatives in the Taylor expansion. To 1st order, this is the same as fitting a plane and estimating the gradient. The quadratic part is tantamount to finding the classical Hessian.

Beaudet considers operators up to 4th order, and operator sizes from 3×3 to 8×8. The only rotationally invariant 1st order operator is the gradient, or rather, more precisely,

the squared magnitude of the gradient, $\nabla f \cdot \nabla f$. The 2nd order operators correspond to the linear invariants of the Hessian matrix,

$$H = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix},$$

as well as the scalar-valued operators $|H\nabla f|$ and $\nabla f H \nabla f$.

Unfortunately, it appears that the author confuses the Hessian with a matrix representation sometimes called the Weingarten map, which is the differential of the Gauss map. The linear invariants of the Weingarten map are the intrinsic curvatures of the surface: the eigenvalues are the principal curvatures, the trace is the mean curvature, and the determinant is the Gaussian curvature. The author, however attributes these properties to the Hessian. This confusion most likely stems from the fact that the two coincide at any critical point of the function f, and it is possible to rotate the 3-dimensional coordinate system of a surface in $\mathbf{R}^3$ so that any given point is a critical point when the surface is being viewed as the graph of a function from $\mathbf{R}^2 \to \mathbf{R}$. This is commonly done in expositions of the subject to simplify formulas. However, since we are in a fixed coordinate system, such a simplification is not possible (without, of course, including the rotation matrices). (See, e.g. [do Carmo 1976].) The differential of the Gauss map, when the surface is given as the graph of a function $f : \mathbf{R}^2 \to \mathbf{R}$, can be written in coordinates x, y as

$$dN = \frac{1}{(1 + f_x^2 + f_y^2)^{3/2}} \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix} \begin{pmatrix} 1 + f_y^2 & -f_x f_y \\ -f_x f_y & 1 + f_x^2 \end{pmatrix},$$

which is easily seen to reduce to the Hessian at a critical point of f.

Beaudet correctly points out that the trace of the Hessian is the Laplacian, but he makes incorrect assertions about the relations between the quantities he derives from the Hessian and various curvatures.

Three 3rd order operators are presented, which are claimed to have significance as line end, curve boundary, and line detectors.

The above terminology and interpretation is mine; he presents these in the more classical language of tensors and coordinates, where his operators are contractions of tensors.

One should note that considerations and techniques very similar to those presented in this paper were described by Prewitt circa 10 years before, though no reference is made to that work.

Evaluation

The experimental results consist in the application of a few of the operators to a single image. Since the notions of line detection and edge detection are very simplistic, there is no effort to use the results of the processing in any way other than to present the magnitude of the operator output. Not surprisingly, this is not very effective. However, more sophisticated processing based on the obtained fit is promising. A potential difficulty may lie in the manner in which the fit is obtained, since polynomial least squares fits tend to produce spurious oscillations.

Despite these shortcomings, the proposal to compute geometrically and analytically significant properties of the image intensity function, using convolutions, is a worthwhile contribution. The thrust, perhaps not made clear by the author, is to derive an understanding of the image intensity function in terms which have precise, well-understood meanings, and which go beyond a few naïvely chosen parameters. As it happens, the error about intrinsic surface properties may be fortuitous, since it may make more sense to consider the Hessian of the intensity function, rather than its surface geometry independent of coordinate system. There is, after all, a special coordinate system in this situation: intensity (the z-axis) is quite different from location (x, y), and so there is no reason to expect that the invariances of rotating the entire 3-dimensional space should be the right ones. It would be interesting to see results of psychophysical studies where the intensity function is changed so that only the Hessian or the Weingarten map, but not both, change.

As presented, this is not a viable edge detection method. However, the idea of local fitting merits further investigation, particularly in regard to deriving differential operators.

Hsu, Mundy, Beaudet 1978

"Web Representation of Image Data"

The authors are interested in using a local quadratic fit to detect image features. A quadratic polynomial is least squares fit to the image data on (presumably uniform) local neighborhoods. The polynomial is regarded as a Taylor series, and the coefficients are interpreted as partial derivatives (see [Beaudet 1978]). The principal axes are identified, and a mesh is constructed over the image by extending straight lines along these special directions until some error threshold is reached, resulting in a new mesh node and repetition of the process. The implementation is based on starting from seed nodes, with special rules for the image periphery, propagating down and right, and merging of nearby nodes. Some nodes of the resulting mesh are labelled according to the "curvatures" and an extremum predicate. Global paths through the mesh are then sought by the use of production rules based on the local labelling to follow arcs. It is not entirely clear how this process works; apparently some kind of relaxation is involved.

Experimental results

Partial results are shown for 1 real and 2 synthetic images of *ca.* 128×128 resolution. Feature finding is only shown for two of these, where a purported ridge is found in a synthetic normal saddle, and some ridges are found in a real picture of scratches. The performance on the real picture is quite poor, although it is hard to isolate the reason. Probably it is a consequence either of the extreme coarseness and irregularity of the mesh, or the localness and ingenuousness of the production rules.

Evaluation

See the review of [Beaudet 1978] for remarks about fitting and differential operators. The same misconception is present here as in [Beaudet 1978] regarding use of the Hessian to define principal curvatures and intrinsic surface properties, rather than the correct expression for the differential of the Gauss map. Consequently, the "principal axes" and "curvatures" the authors find correspond to the conventional usage of those terms only at stationary points of the image intensity function. However (see review of [Beaudet 1978]), these objects may actually be more meaningful for image analysis than the geometric invariants. E.g., [Canny 1983] uses essentially the same parameters in his directional operators.

The construction of the mesh is a good idea insofar as a coordinate system based on principal directions is found. However, the mesh is far too coarse, and the method of its construction leads to a topology which may not have much to do with the underlying structure. The authors apparently wanted a graph structure to propagate their production rules on, but unless they have bugs, what they got was more or less a mess. The production rule technique is not very well explained, hence difficult to evaluate, but the impression one gets is that it is somewhat inflexible, e.g. putting limits on rotation of principal direction. It is not clear, e.g. how the production system performs a function separate from the mesh generation itself, where error criteria are also imposed. It may be that using a finer mesh would provide much improved results.

A second problem is that no analysis is given regarding noise behavior. A big question is the behavior of the mesh generation in the presence of noise.

Dreschler and Nagel 1981a, Dreschler and Nagel 1981b

"Volumetric Model and 3D-Trajectory of a Moving Car Derived from Monocular TV-Frame Sequences of a Street Scene"

The authors are primarily interested in tracking objects in a sequence of successive static frames. They seek point features which are expected to be stable from frame to frame, settling on extremal points of the Gaussian curvature of the intensity function. The computation of the curvature is performed via "principal curvatures" using the operators of Beaudet (which in fact compute something other than principal curvatures: see review of [Beaudet 1978]).

The authors are motivated by seeking local extrema of Gaussian curvature. However, they found that such extrema occur at knees of edges (cliffs in the intensity function) in an unstable manner, as a consequence of local noise and small variations. Therefore, a more involved predicate is used. Viz., pairs of nearby points are found which are a maximum and a minimum of Gaussian curvature. Along the line joining these points, that point having the steepest slope of intensity (i.e. directional derivative) is selected as the feature point, subject to the following 2 criteria. First, it is asserted that exactly 1 principal curvature must change sign along the line in question (this is true only if the extrema of Gaussian curvature are of opposite sign, which is implicitly assumed), hence it is required that the principal direction corresponding to the principal curvature which is changing sign be roughly parallel to the line in question. This assumes that the extrema of the Gaussian curvature should be joined by principal curves, a proposition whose truth is by no means self-evident. Secondly, the intensity value at the maximum must be greater than that at the minimum. This is for the case that the high intensity area is convex at the corner. Since the reverse case obtains by turning the surface upside down, which does not change the Gaussian curvature anywhere, the opposite condition must be true when the low intensity area is convex, so without other information about the context of the extrema, this seems to be a vacuous condition. Also, an *ad hoc* maximum separation of 4 pixels is required for pairs of extrema to be linked. Obviously, this is a requirement that the corner be quite sharp at the resolution of the image.

Both 5×5 and 3×3 operators are used: the 5×5 for good noise behavior, and the 3×3

for better resolution in places selected by the 5×5. The operators used are the ones presented in [Beaudet 1978]. Consequently, the present authors are victims of an incorrect definition of Gaussian curvature (see review of [Beaudet 1978]) and principal directions. However, it is extrema of Gaussian curvature which are of interest. The relation between these and what is actually (erroneously) used is algebraically complicated, and we do not attempt to analyze it, but these parameters may be just as meaningful for images as the geometric ones. Furthermore, there is already a heuristic element to locating the points of interest. Therefore, it doesn't seem likely that the use of the correct values of the Gaussian would change the performance significantly. To get a better understanding of the situation, one should in fact analyze the behavior of these parameters in the light of what is known about the image irradiance equation.

Experimental results

The results displayed seem to be fairly good. Of course, there are a number of other elements of the system we are not considering here, e.g. the method of tracking, so that it is difficult to say how reliably the features selected represented intrinsic features of objects or even of the intensity function.

Evaluation

The present work is best regarded as a corner detector. As such, it is not adequate for performing segmentation. As far as its usefulness for matching images is concerned, one would have to analyze to what degree extrema of Gaussian curvature are intrinsic features of the object geometry, rather than the intensity surface geometry. There are 2 components to such a study: the effects of perspective transformation, and the effects of photometric laws. An initial approach could consider these components separately, i.e. constant light with moving observer, and fixed observer with moving light source. Since the features used are piecewise smooth functions of the parameters of motion and lighting, one can expect that they will trace out piecewise smooth curves as those parameters are

varied; and hence they can be tracked. Whether they are good things to track is another question. Consider the extreme case of a moving flat mirror, moving in its own plane, and reflecting a light source. This isn't a completely ridiculous case, since it is a limiting case of what can happen with specularity, which in turn is a matter of degree for the reflectance function. The point to note is that the feature associated with the specularity will behave as a function of the location of the light source rather than as a function of the motion of the object reflecting it. The moral is that the behavior of a feature can be highly decoupled from that of the object whose surface creates it. A less extreme example to ponder is studied by [Koenderink and van Doorn 1980], who show that the extrema of the image intensity stay near parabolic lines of the object surface (but move along them).

The relevance to image segmentation is this. Principal curvatures, principal directions, and principal curves are useful features of the image intensity function. They define a local geometry, and notably a local orthogonal coordinate system which is a natural coordinate system in the vicinity of edges. Predicates based on observation of the behavior of principal curves seem good candidates for edge detection and hence segmentation. This work at least shows that such features have some stability in the presence of noise and deformation.

Haralick 1980

"Edge and Region Analysis for Digital Image Data"

The view taken here is that edges and regions can be viewed as places where there are large or small differences, resp., in some parameters. In this light, the old method, i.e. looking for perfect step edges, amounts to fitting a piecewise constant function to the image intensity. The new method which the author puts forth, is to do a piecewise *linear* fit, i.e. to fit planes (or *facets*). The work is purely theoretical in that real images are not considered.

The central feature of the analysis is to perform a least squares fit of a plane to the data.

The author provides a nice analysis of noise for this problem. The critical question is whether 2 planar patches are actually part of the same plane: the edge null hypothesis. To resolve this question, he uses the F-test on a χ^2 distribution derived from the error.

More specifically, the way this is used is as follows. Each point p of the picture is assigned a neighborhood $p \mapsto U_p$ which is the one supporting the best fit among all U_p^i containing p. I.e., of all neighborhoods U_p^i such that $p \in U_p^i$, let U_p be such that $\epsilon(U_p)$ is minimum.

Edge and region detection are then based on an F-test of the parameters associated with the optimal neighborhoods for adjacent pixels, followed by thinning. Even neglecting the piecewise planarity assumption, this adjacent-F-test is probably too simple minded.

The technique can be summarized as follows:

edge detection method:

each pixel has a best-fit neighborhood with parameters of fit.

edgeness = F statistic that adjacent pixels' fits come from same plane.

compute for vertical, horizontal adjacencies for vertical, horizontal edges.

find maxima by non-maximum suppression.

region growing method:

group adjacent pixels if same best fit neighborhood plane hypothesis cannot be rejected.

The hypothesis testing is based on the relation between parameter differences and errors of fit. If the local is relatively poorer, greater parameter differences are tolerated for region merging. In this sense, the region merging is adaptive. However, no analysis is presented describing how this method would behave for large regions or long edges. Also, no attention is given to the problem of determining whether local edges are part of a larger edge.

The author includes a quick but nice review of some related literature. For example, he shows that the "Roberts cross" operator [Roberts 1963] computes the magnitude of the

gradient of a linear fit (although this is all but explicitly stated in [Prewitt 1970]).

Unfortunately, the paper includes no experimental results or consideration of real images.

Evaluation

The idea of fitting regions and looking at the parameters is a good one. Statistical analysis is good, too. However, the piecewise planar hypothesis is not sophisticated enough. On the other hand, the statistics becomes more complicated for more complicated fits. In the form proposed, this method is not likely to be noticeably better than other local methods. The extended edge and region part is rather ad hoc—not based on a sound analysis. This paper can be recommended as a good introduction to the use of statistics and fitting, despite some ambiguities.

Haralick 1981, Haralick 1982, Haralick 1984

"Digital Step Edges from Zero Crossing of Second Directional Derivatives"

The essential feature of the technique proposed by the author is fitting the image intensity function by a polynomial.

He first intuitively defines edges as discontinuities in brightness value or its "derivative." But then he notes that for this to make sense, the discrete picture must be thought of as samples of a function on a continuum. To obtain such a function from the data, he does polynomial approximation using discrete orthogonal polynomials. "Discrete orthogonal" means orthogonal with respect to the "inner product"

$$(f, g) = \sum_{p \in P} f(p)g(p)$$

where P is some finite set of points, though this is not explicitly stated. It is not a true inner product because it can happen that $(f, f) = 0$ with $f \neq 0$. (see e.g., [DeBoor

8]). Regrettably, he provides no references: there is, after all, a rather large literature taining to fitting polynomials.

hin this context, he can define what is meant by "edge." In [Haralick 1981], this is ned as a place where the "direction isotropic magnitudes" of the 1st or 2nd partials of fitted function exceed some threshold. However, he requires the assumption that the ivatives of the underlying function are uniformly bounded except at discontinuities that the high estimated values can be attributed to discontinuities). This is not very listic, and in [Haralick 1982 and Haralick 1984] it is replaced by a definition of "edge" a zero crossing of the 2nd directional derivative in the gradient direction (see review [Canny 1983] for a more detailed definition of this entity), i.e. a maximum of the dient. While looking at the parameters of a fitted function is a good approach, this still too local a criterion, and too simplistic a structural representation, so that most the benefits of surface fitting are lost, as demonstrated by [Canny 1983]. In [Haralick, tson, Laffey 1983], he improves this considerably, expanding to the derivation of the litative structure of the function.

imposes 1-dimensional symmetry on the index sets of the polynomials, i.e. the points which they are defined must be symmetric about the origin. For 2-dimensional basis ctions, he uses the tensor product of his 1-dimensional set. He then shows how to by the usual method of projection onto the orthonormal basis. A further section of ralick 1981] is devoted to showing that $D_x^2 + D_y^2$ and $D_{xx}^2 + D_{yy}^2$ are rotationally ariant differential operators.

luation

e idea of fitting a function to the intensity data as a first step in edge finding is good, hough the definition of "edge" is somewhat simple-minded. E.g., the 1st derivative erion will result in edges being found in regions of smooth shading. Unfortunately, papers do not address issues associated with the fitting problem. E.g., polynomial

done, which Canny tackles with numerical methods, yielding a family of optimal olution kernels, parametrized by K, the mean separation of maxima normalized by upport interval. Qualitatively, this family ranges in appearance from a smoothed ence of boxes for small values of K to a derivative of Gaussian for large values of

he same time, he develops another measure of multiple response, a local measure ı by

$$\frac{|f'(0)|}{\|f''\|}$$

ct a signal-independent constraint, a proportionality can be required between this sure and the false positive (detection) measure, since they are both normally dis- ıted; i.e. one can require

$$\frac{|f'(0)|}{\|f''\|} = k\frac{(f, \tilde{u}_{-1})}{\|f\|}$$

rms of previously defined quantities, this can be written as

$$\Lambda K W = k\Sigma$$

ough Canny seeks an f for which $k = 1$, the best he is able to do is $k = .58$, which ıt too surprising, since at this point the constraints are no longer all independent. value is achieved for one of the larger values of K. The f thus arrived at is well oximated by a derivative of Gaussian, which is desirable for ease of computation, icularly in a 2-dimensional version. However, aside from computational considera- ı, it is not entirely clear that this is a necessary choice. Canny does not make it that one necessarily wants $k = 1$, and for that matter, the argument leading to the response measure used in defining k is less convincing than one would like. It would ı that where a 1st derivative of noise response is used, that a 2nd derivative (possibly aged over some neighborhood) should be used. In any case, he computes that the

raw (i.e. signal-dependant) SNR and localization terms do not cancel when these terms are multiplied, but result in a coefficient of A^2/n_0^2, i.e. if these terms are not dropped in defining the detection and localization criteria, the resulting product would be

$$\frac{A^2}{n_0^2}\Sigma\Lambda$$

The problem of finding an f to optimize the composite measure is solved as a variational problem, making an assumption of finite extent and thereby using a tractable formulation. The set of admissible functions is taken to be C^0, which may be slightly inconsistent, since it would seem that f must be at least C^1 to conclude that the maximum of $f * I$ will be achieved where the derivative is 0, which was used in the derivation of the optimization measure. Solving the variational problem leads to an expression depending on a parameter (for normalized f). It turns out that the parameter can be increased without bound, leading to ever better f's, and, in fact the limit of the f's is a difference of boxes (not in the admissible space), which, not too surprisingly, is the Wiener filter, giving infinitely good localization, and the best SNR.

Multiple response criterion and optimizing for all criteria

Now, if f is a difference of boxes, $f * I$ is no longer smooth, so the derivative method of finding maxima is called into question. But what is more important, as Canny notes, the maxima will be essentially as noisy as the noise. This observation leads to an excellent way of imposing a smoothness constraint on f. Namely, one can couple the requirement that maxima of $f * I$ be sufficiently isolated (in the mean) with a formula giving the separation value to arrive at a smoothness constraint on f of the form

$$\frac{1}{2\pi} \cdot \text{mean separation} = \frac{\|f'\|}{\|f''\|} = KW$$

where W is the support width, and K the parameter which sets the constraint in units of W. This leads to a complicated algebraic problem once the variational work has

lo one can write

$$-\frac{(f' * n)(x)}{Af'(0)} = x + \frac{1}{f'(0)} \cdot (\text{higher order terms})$$

Taking root mean square expectation values, one gets

$$\frac{n_0\|f'\|}{A|f'(0)|} = E^{1/2}(x^2 + ax^4 + \text{higher order terms})$$

The right side is taken as an approximation to the standard deviation of x, the solved-for location. The localization measure is taken as the reciprocal of the left side, i.e.

$$\Lambda = \frac{|f'(0)|}{\|f'\|}$$

Λ, then, should also be maximized.

Optimizing sensitivity and localization

Canny chooses to optimize over the composite measure

$$\Sigma \cdot \Lambda = \frac{(f, \tilde{u}_{-1})|f'(0)|}{\|f\| \cdot \|f'\|}$$

based on the observation that this is a scale invariant quantity, i.e. its value is the same for $f(x)$ as for $f(\alpha x)$. While this seems to be an interesting property, the only argument presented in its favor is that the resulting measure depends only on the "shape" of f. It would be interesting to put this on some stronger footing. E.g., the noise is scale invariant, and so is the step (when considered as a function, though not as a distribution), so there is a symmetry argument for scale invariance. On the other hand, $\Sigma^2 + \Lambda^2$ or $(\Sigma + \Lambda)^2$ (where Σ is redefined to be always nonnegative) also seem like reasonable candidates for measures to be optimized. Incidentally, one should note that the A/n_0 coefficients in the

where $(\cdot,\cdot)$ and $\|\cdot\|$ are the L^2 inner product and norm, respectively, $\tilde{u}_{-1}$ is defined by $\tilde{u}_{-1}(t) = u_{-1}(-t)$, and n_0 is the RMS noise. A noise figure Σ for the operator f can then be defined by

$$\Sigma = \frac{(f, \tilde{u}_{-1})}{\|f\|}$$

Part of the optimization, then, is to maximize Σ.

Localization criterion

The localization is given by the location of the maximum of $f * I$. Canny equates finding this maximum with solving

$$(f * I)'(x) = 0$$

This amounts to a smoothness assumption on $f * I$, which unfortunately is not explicitly stated, which makes it unclear what function spaces are involved at various stages. Since $I = Au_{-1} + n$, this is the same as solving

$$(f * Au_{-1})'(x) + (f * n)'(x) = 0$$

$(f * Au_{-1})' = Af * u_0 = Af$, and $(f * n)' = f' * n$, so we have

$$Af(x) + (f' * n)(x) = 0$$

I.e., we want to solve

$$Af(x) = -(f' * n)(x)$$

Canny approaches this problem by first observing that f should be an odd function, then linearizing the problem as follows. Near 0, f can be approximated by its Taylor expansion, so we can write

$$Af(x) = A[f'(0)x + \text{higher order terms}] = -(f' * n)(x)$$

of both variational and numerical methods. The operator is extended to a directional family for 2 dimensions. He uses an adaptive thresholding technique and a noise based scale selection technique to finally output a very clean set of linked edges.

The 1-dimensional problem

The 1-dimensional problem which he poses is this. Assume that the data consists of some step function in white Gaussian noise, i.e. the data is given by

$$I(t) = Au_{-1}(t) + n(t)$$

where A is a real constant, $u_{-1}(t)$ is the unit step function, and $n(t)$ is the noise process.

Assume further that edge detection proceeds by finding the maxima of $f * I$, for some convolution kernel, f. The problem is to find the best f subject to the following performance criteria:

1) Good detection: low false negative, false positive. Equivalent to maximizing S/N (signal to noise ratio).

2) Good localization.

3) 1 edge yields only 1 response.

Sensitivity criterion

The signal to noise ratio is given by

$$\text{SNR} = \frac{\text{signal response}}{\text{RMS noise response}} = \frac{A\int_{-\infty}^{0} f(x)dx}{n_0(\int_{-\infty}^{\infty} f^2(x)dx)^{1/2}}$$

which we can write more compactly as

$$\text{SNR} = \frac{A(f, \tilde{u}_{-1})}{n_0\|f\|}$$

[Shanmugam, Dickey, Green 1979]. In any case, as a 2nd derivative operator it is esthetically pleasing because of smoothness and the Fourier domain symmetry (i.e. the Gaussian is an eigenfunction, or Gaussians in general are an invariant subspace of the Fourier transform). Zero crossings are a useful way to locate edges, but none of the "mathematical" or heuristic arguments presented about them here are convincing.

One must regard the assumptions and techniques of this work as tentative and experimental, rather than as a well founded theoretical or practical system. The ideas are based on intuition, perhaps good intuition, but lacking better justification must be regarded as only intuitive. The professed purpose is an explication of human vision. Unfortunately, so little is known about human vision (e.g. there is no viable theory of how any but the most rudimentary information is coded or utilized), that one cannot draw any conclusions about the validity of any theory purporting to explain human vision, and in any case it is not our purpose to do so here. For example, it is clear that there are on-center off-surround receptive fields with a response qualitatively like the DOG. But one can approximate the same data with polynomials, Bessel functions, or your own favorite. The important thing is the qualitative feature of smoothly varying on-center off-surround response. The DOG may be computationally convenient, which is reason enough for its use, but the type of convenience does not translate into cost for a living system, without further analysis. It is fair to say that the theory presented here is not obviously ruled out, but neither is it clearly the best or only possibility. As far as its being a theory of visual information or an engineering design, one can only say that it is an interesting and provoking hypothesis, but not inexorable or proven.

Canny 1983

"Finding Edges and Lines in Images"

In this work, Canny begins by posing a local 1-dimensional edge detection problem as an optimization problem over the set of convolution operators, which he solves by the use

$[\partial^2 f/\partial x_i \partial x_j]$. Note that $D_v^2 f = D_v(D_v f)$, where D_v is the directional derivative in the v direction). Unfortunately, they seem to be mainly thinking about the special case where $\partial f/\partial y = 0$; they develop some strange ideas about the conditions and ramifications of their maximum slope concern. They state and claim to prove a theorem to the effect that the condition (confusingly stated) obtains if and only if $\partial f/\partial y = constant$. They are somewhat careless in the proof of the *if* part, failing to explicitly consider the slope. In fact, what they are trying to show, in their notation, is that $\cos^3\theta \cdot f_{xxx}$ attains a strict maximum for $\theta = 0$. This will be true if f is 3 times differentiable and $f_{xxx} \neq 0$. However, the authors have only assumed that $f \in C^2$, and they neglect the possibility that f_{xxx} may vanish. The *only if* part and its "proof" are omitted from the Proc. R. Soc. version of the paper, and wisely so, for they are erroneous; the purported proof shows only that $D_v^2 f$ may not be 0 for $v \neq \hat{x}$. See [Canny 1983] for a more coherent use of 2nd derivative, and the review of [Canny 1983] for more discussion.

The authors assume that coincident zero crossings from a set of contiguous channels imply a real edge and conversely. This so-called *spatial coincidence assumption* is not well-supported by any argument. (E.g. see [Canny 1983] for pictorial counterexamples.) The only situation for which it really makes sense is that of a very sharp edge between fairly large constant areas. Otherwise, it seems perfectly reasonable to believe that the edge will be visible at only 1 scale, while smaller scales will have inadequate sensitivity, i.e. their zero crossings will be essentially random, and larger scales may include other features, so their zero crossings will depend (arbitrarily) on those features as well.

Evaluation

The paper presents no convincing arguments that the $\nabla^2 G$ or DOG operator is optimal or otherwise privileged in this context. However, for the purpose of step edge detection a particular kind of optimality under some conditions has been shown elsewhere

matched, yielding a simple result in the case of a step edge. In that case, the approximate filter has an impulse response which is the 2nd derivative of a Gaussian (see review of [Shanmugam, Dickey, Green 1979] for details). For the ranges that the approximations are valid (see review of [Shanmugam, Dickey, Green 1979]), this vindicates the use of the Laplacian of the Gaussian by Marr and Hildreth, but only for the specific type of matched filtering of *step edges* studied by [Shanmugam, Dickey, Green 1979], though [Marr and Hildreth 1979] makes no mention of the type of analysis in [Shanmugam, Dickey, Green 1979], basing the use of the Gaussian on the more nebulous grounds mentioned above.

The authors are interested in finding points of maximum directional derivative as edge locations, and they choose to locate these as zero crossings of a 2nd derivative. Based on cost considerations they opt for an isotropic 2nd derivative operator, the Laplacian ∇^2 (the only such), and wish to compute $\nabla^2(G * f)$, where G is the Gaussian. Since $\nabla^2(G * f) = (\nabla^2 G) * f$, they want to convolve with $\nabla^2 G$, which they approximate as a difference of Gaussians (DOG).

Logan's theorem (reconstructibility of analytic 1 octave bandpass signals from their zero crossings) is invoked to help justify use of zero crossings. However, the theorem is applicable only for 1 dimension, and the signals involved here have a bandpass of nearly 2 octaves. An argument is made that slope information may be adequate to bridge the gap (in analogy to the situation for the sampling theorem). On the other hand, there is no reason why reconstructibility should be a criterion, since there is never any requirement that an image *understanding* system should be able to reconstruct the input signal.

It appears that the authors are concerned that the zero crossing direction be perpendicular to the direction of "maximum slope of the directional derivative." Apparently, what this is supposed to mean is that $\nabla D_v^2 f$ should be collinear with v (where D_v^2 is the 2nd directional derivative in the v direction, the second derivative of a section of f taken along a line in the v direction, which can be written as $v^T H v$, where H is the Hessian matrix

lead to the frequency domain. If one regards "scale" as referring to rate of change, then normalizing a bandlimited function bounds the derivative, but the converse need not be true. Thus, bandlimiting can be regarded as one way to limit scale in some sense. However, no arguments are presented to bolster the desire to consider the frequency domain. The reason that frequency domain methods work in engineering is the fact that exponentials are eigenfunctions of linear translation invariant operators, so one can use superposition to combine the effects of various bandpasses. Related is the convenient fact that convolutions are mapped to multiplications. The work under consideration uses exclusively linear methods, but does not present such an argument. On the other hand, if one uses nonlinear methods, there is no such justification. (See the review of [Shanmugam, Dickey, Green 1979] for another argument supporting bandlimiting.)

The authors argue further that the conflicting requirements of space- and band-limiting are optimally reconciled by minimizing the space-bandwidth product. For the appropriate definition of these terms, it is well known that the Gaussian (e^{-kx^2}, for the right k) achieves the minimum, so the authors conclude that the filters they want are Gaussian. Unfortunately, even if one accepts the doctrine of band-limiting, it is by no means clear that the Gaussian is optimal. In the first place, the Gaussian is neither strictly band-limited nor strictly space-limited. When one, say, bandlimits by truncation, it is no longer optimal. If one requires a strictly band-limited or space-limited function, i.e. one which is 0 outside of a given interval in either the spatial or frequency domain, the work of [Slepian and Pollak 1961, Landau and Pollak 1961, Landau and Pollak 1962] and [Shanmugam, Dickey, Green 1979] shows that the optimal filter has a transfer function which is essentially a prolate spheroidal wave function divided by the transform of the waveform to be matched, where optimality is defined in terms of concentrating energy in a spatial interval, rather than minimizing space-bandwidth product. Under some conditions, the prolate spheroidal wave functions can be approximated by functions related to the Gaussian. However, the optimal filter still depends on the function to be

be strictly space-limited. This is the case if one convolves a mask with the image. If one were interested in concentrating energy in some band, then the problem would be the dual of the one considered in the reviewed paper, viz. to find the optimal space- (or time-) limited filter for concentrating its energy in some frequency band. With the duality of the Fourier transform, the solution is essentially the same. Now an argument can be made for considering the frequency domain based on noise considerations, for as the authors show, the signal to noise ratio is a function of the space-bandwidth product. Since white Gaussian noise has constant spectral power density, the frequency domain is a natural setting for its analysis. Unfortunately, for good present-day images, the true noise is of the same order as the digitization noise, and most of the "noise" really comes from real variations in the image, i.e. from the fact that the image is not in the space of ideal features. It is not clear whether this type of "noise" can properly be regarded as white and Gaussian; e.g., it is not perfectly uncorrelated.

On the other hand, it would indeed be satisfying to learn that bandlimiting *is* required for some strong inherent reason, so the prolate spheroidal wave functions are worth experimenting with, and should at least be kept in mind.

Marr and Hildreth 1979

"Theory of Edge Detection"

The authors are concerned with finding a smoothing filter which will analyze the visual input into a number of channels related to physical scale.

They argue that such a filter should operate over a subrange of scales—not over all scales possible in the image. Furthermore, it should be spatially localized. From this they infer the (contradictory) requirements that the filter be both band-limited and space-limited. Although space limiting clearly follows from the localization requirement, the band-limiting conclusion is on shaky ground, since the idea of "scale" does not inexorably

where Ω is the bandwidth (i.e. the signal is nonzero only when $\omega \in (-\Omega, \Omega)$), and the energy is to be concentrated in the spatial interval $(-I, I)$ (note this is a slightly different use of I than in [Shanmugam, Dickey, Green 1979, Lunscher 1983]). These conditions say that the approximation is valid for large space-bandwidth products, and under those conditions it is valid only away from the band limits. If those conditions are violated, e.g. by requiring better localization, then the prolate spheroidal wave function which is the solution no longer looks like a Gaussian. This is similar to the localization results found by [Canny 1983].

Blurred edges are modelled as the difference of exponentials to obtain a symmetrical sigmoid function (only once continuously differentiable, though). They show that if the resolution interval I is larger than the blur width (defined by the 90% points), then the filter is still a good approximation to optimal in an appropriate sense.

A Gaussian noise analysis is also presented, showing that S/N improves with increasing space-bandwidth product, e.g. coarser resolution, not a very surprising result in view of many others to the same effect. An expression for S/N is given.

The experimental results are not very impressive when compared to nonlinear edge detectors (e.g. after thresholding), but they show a clear improvement over other standard linear filters, e.g. high pass, Laplacian.

Evaluation

This is not a direct method of detecting edges, but rather should be regarded either as an enhancement method, or, more importantly, as a precise approach that could be taken in finding an optimum filter to reconcile space- and band-limiting.

If one must do computations in the frequency domain, then the filter used must be strictly band-limited. But there is no persuasive argument for using the frequency domain. If one does computations in the spatial (or time) domain, then of course the filter must

where K_1, K_2 are simple functions of Ω, I. When the ideal input is a step edge, this reduces to

$$H(\omega) = K_1 \omega^2 e^{-K_2 \omega^2}$$

The authors allude to work by Streifer [Streifer 1965b] showing that "the error is not prohibitive even when Slepian's constraints are violated." [Lunscher 1983] has pointed out a dimensional error in the exponent above, and uses asymptotic expansions of [Streifer 1965a, Streifer 1965b] to arrive at a K_2 of the correct dimensions to assure a scale-invariant response.

The optimality of Gaussians

What does this say about the optimality of Gaussians? Since Gaussians minimize the space-bandwidth product for functions of infinite (frequency) extent, one would expect that the imposition of a finite extent constraint would lead to a result which approached a Gaussian asymptotically. The question then becomes whether the conditions for the asymptotic approximation are applicable in a particular situation. For example, if one starts with a Gaussian which is approximately band-limited, say 99% of its energy is within $(-\Omega, \Omega)$, then that Gaussian has a particular spatial extent, too, parametrized by its standard deviation, so 99% of its energy is in the spatial interval $(-I, I)$, where I is the appropriate multiple of σ. Now if we are demanding that the function we are interested in must concentrate its energy in the interval $(-.01 \cdot I, .01 \cdot I)$, then clearly the Gaussian will not be a very good approximation.

For the scale-invariant version of the prolate spheroidal approximation due to [Lunscher 1983], the domain of validity is defined by

$$\Omega I \gg 1$$
$$|\omega| < \frac{\Omega}{(\Omega I)^{\frac{1}{2}}}$$

Following [Slepian and Pollak 1961, Landau and Pollak 1961, Landau and Pollak 1962] they decompose in terms of prolate spheroidal wave functions, and show that the optimal filter output is ψ_1, the order 1 prolate spheroidal wave function, with the space-bandwidth parameter dependent on the space and bandwidth cutoffs required. This method of analysis allows the bandwidth and space cutoff to be chosen *independently*, unlike the situation with a Gaussian. This constitutes a more realistic treatment of the type of optimality sought in [Marr and Hildreth 1979], yielding functions other than Gaussians, although under certain ranges of parameters the Gaussian is a good approximation. Specifically, the transfer function of the optimal filter is given by

$$H(\omega) = \begin{cases} K\frac{\psi_1(\Omega I, \frac{\omega I}{\Omega})}{iF(\omega)}, & |\omega| < \Omega \\ 0 \quad , & |\omega| \geq \Omega \end{cases}$$

where K is a real constant, ψ_1 is the 1st order prolate spheroidal wave function, Ω is the half bandwidth (i.e. the signal is nonzero only when $\omega \in (-\Omega, \Omega)$), and the energy is to be concentrated in the spatial interval $(-I, I)$ (note this is a slightly different use of I than in [Shanmugam, Dickey, Green 1979, Lunscher 1983]), and $F(\omega)$ is the Fourier transform of the ideal input. The only information used about the input and filter to derive this formula is the fact that they are odd and even functions, resp. There is no particular justification for requiring the filter to be even (it gives a neater result) except that it allows ready generalization to a rotationally invariant 2-dimensional operator simply by making the value depend only on distance from the origin, i.e. by rotating the 1-dimensional operator. [Canny 1983] regards this as a rather unfortunate assumption, since in his analysis directional operators provide better sensitivity, and he shows that dropping the assumption leads to an operator very much like the one he proposes.

Using an approximation of Slepian [Slepian 1965], the optimal filter within the bandpass is approximated as

$$H(\omega) = \frac{K_1 \omega e^{-K_2 \omega^2}}{iF(\omega)}$$

least squares fits (which are being proposed) are notorious for being badly behaved—they tend to have extra wiggles. One would expect that such functions would not be very good ones to use if one wanted to look at derivatives. One might prefer to use Fourier interpolation, B-splines, Fourier splines, or some other appropriately well-behaved set of functions. No mention is made of his previous idea of looking at discontinuities of *parameters* of fit between adjacent regions. Nevertheless, some kind of fitting process seems to be in order to use global information for local features (in this case the global fit yields the local derivative). The noise performance issue is postponed in [Haralick 1981], but treated thoroughly in [Haralick 1984].

These "edge" detectors are a beginning based on surface fitting. The particular predicates involved are not adequate, though, and therefore cannot be expected to give outstanding performance (see [Canny 1983] for one discussion of performance). Improvements can be expected when the type of qualitative information used in [Haralick, Watson, Laffey 1983] is brought to bear on finding edges.

Optimal Filters

Shanmugam, Dickey, Green 1979

"An Optimal Frequency Domain Filter for Edge Detection in Digital Images"

The authors consider the 1-dimensional edge detection problem, with the proviso that "symmetries appropriate to the 2-dimensional problem are retained." Their goal is to obtain a frequency domain filter to concentrate maximal energy near an edge. The model for an input edge is the unit step.

More particularly, the authors require a strictly bandlimited filter (i.e. a filter whose Fourier transform has its support on an interval surrounding the origin), and they seek to maximize the power in some interval around the origin in the space domain for the filter output response to a unit step.

Gaussian approximation has a performance measure $\Sigma\Lambda$ which is about 20% worse than the optimum operator.

The 2-dimensional problem

Canny does not consider the 2-dimensional optimization problem *de novo*, but rather starts from the point he reached with the 1-dimensional problem, which is the derivative-of-Gaussian operator. The approach is to use an operator of the form $h(x,y) = f(x) \cdot g(y)$, for various orientations of the orthogonal coordinates x, y. Then f is to be the (approximate) optimal 1-dimensional operator, and g must be determined. By reasoning similar to that involved in finding f, he notes that g should be smooth, i.e. a smooth window function, and he notes that the Gaussian he chooses is a good approximation to standard windowing functions. First, the edge orientation is estimated from the gradient of the smoothed image, i.e. from

$$\nabla(G * I)$$

where G is a rotationally symmetric Gaussian. Then the location of the edge is determined by finding the zero crossings of an operator which computes $D_v^2 G * I$, the 2nd directional derivative in the v direction, where v is approximately given from the gradient estimate. This is what it would seem [Marr and Hildreth 1979] were really after. Notice that this can be realized as a single operator (i.e. it is not a directional family) since one seeks the zeroes of

$$D_{\nabla S} S$$

where $S = G * I$.

A compact description of the 2nd derivative operator is as follows. The 2nd derivative for a function f of 2 variables can be thought of as a matrix, known as the *Hessian* matrix, given by $H = [\partial^2 f / \partial x_i \partial x_j]$. The 2nd directional derivative in the v direction is then

given by $D_v^2 f = v^T H v$ When $v = \nabla f$, this can be written

$$D_{\nabla f}^2 f = \begin{pmatrix} f_x & f_y \end{pmatrix} \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix} \begin{pmatrix} f_x \\ f_y \end{pmatrix}$$

which expands to

$$D_{\nabla f}^2 f = f_x^2 f_{xx} + 2 f_x f_y f_{xy} + f_y^2 f_{yy}$$

Normalizing by $|\nabla f|$ does not alter the zeroes of this quantity.

Canny makes a useful observation in comparing this directional derivative operator to the Laplacian, $\nabla^2(G * I)$, which is worthwhile repeating here. Consider a coordinate system with the x-axis aligned with the gradient direction (at the point of interest). In this coordinate system, the directional derivative has a contribution only from f_{xx}, since $f_y = 0$, since the y direction is orthogonal to the gradient, which is as it should be for a directional derivative. The Laplacian, on the other hand, also is invariant under rotation, but does not depend on the gradient, hence it has a contribution from the 2nd derivative in the "uninteresting" y-direction, which leads to nothing but a noise contribution. Actually, this is a little more subtle than may appear at first glance. If one assumes an ideal edge as signal embedded in noise, then the signal is completely constant in the y direction; hence all y derivatives will be 0, so in fact the directional derivative gives the same answer as the Laplacian (modulo a first order coefficient which would be normalized away), *for the signal response alone.* But with noise, the Laplacian will respond in the y direction, while the directional derivative will not.

From this theme (the second directional derivative of Gaussian convolution), Canny proceeds to develop a number of variations: multiple widths, "feature synthesis," elongated operators, and lateral inhibition.

Multiple widths are required since sensitivity increases with size of support, while localization degrades. Canny's approach is to use the smallest operator with sensitivity adequate

to provide a given error probability. This requires estimating the noise, which he does by convolving a filter with the edge detector output. Under the assumption that the signal is an ideal step and the step size is much larger than the noise amplitude (low noise), he finds that the optimal filter for this is the 2nd derivative of a delta function, and smoothing the response gives the 2nd derivative of a Gaussian. This he approximates with a difference of Gaussians, which is less sensitive to the accuracy of the edge location estimate, with coefficients chosen so as to make it orthogonal to the step response. Of course measuring noise involves a model of the image, in this case an ideal step, so the noise measurement is also a measurement of the deviation from the model. Nevertheless, since it is also a measure of the fit of the model, it is still useful as a confidence measure.

It is a fact of life that images are not composed of ideal step edges. Consequently, operators of different sizes with adequate S/N centered at the same point may be responding to different aspects of the image function. The simplest example is a diffuse edge superimposed on a sharp one (possibly at a different orientation). In general, the single number that a filter gives at a point does not convey a great deal of information about the structure of the image in a neighborhood of that point. In particular, the response of an edge operator based on an assumption of ideal edges gives very little information about the shape of actual edge candidates. Canny's approach to this problem, "feature synthesis," is reminiscent of the Gram-Schmidt orthogonalization procedure. The idea is that, starting with the response from the smallest significant operator, he estimates what the response from the next largest would be if a step edge were responsible. If there is a large enough disparity with the observed response it is deemed to come from something else. This has the effect of enlarging the feature space.

Elongating a mask along the edge direction is another way to increase the support, hence the sensitivity, but since there is no scale change, the localization improves (under the ideal straight edge assumption). Canny's elongated operator is essentially the sum of Gaussians taken along an interval, resulting in a mesa shape with Gaussian fall-off.

A common problem with local operators that respond to the gradient is that they respond to slow changes as well as abrupt ones. This can be regarded, again, as a symptom of an inadequate feature space (the 1-dimensional "edgeness" number, essentially a projection to a 1-dimensional space). One remedy for this problem is to introduce a preprocessing step, ***lateral inhibition***, which sends offending subspaces to 0. In the context of the Taylor expansion, the first offending subspace is the constant term, but this is already taken care of as long as the operator has 0 average value, e.g. if it is an odd function. The next problem is the 1st order term, which is an example of a "smooth gradient." This can be removed by some 2nd order operation, e.g. 2nd derivative. Canny uses a difference of 1st derivative of Gaussians of different widths, weighted to send linear functions to 0, i.e. roughly the difference of adjacent channels of the optimal operator. Like other lateral inhibition methods, this degrades the performance, in this case by about 30%. It's not too hard to see what the problem is here. First of all, the operator f was chosen to maximize

$$h = \int f \cdot u_{-1} = (f, u_{-1})$$

Without any constraints (and an appropriate measure), this is achieved for $f = u_{-1}$. With the extra constraints, we can think of it as finding the dual vector of u_{-1}, or we can put everything into the measure, in this case Gaussian measure. Now to compute the u_{-1}-ness of I, we compute (f, I), i.e. we apply the distribution f to I, or equivalently, look at the projection on the u_{-1}-axis. Unfortunately, it turns out that $(f, t) \neq 0$ (where t stands for the identity function on the line). I.e. $u_{-1}(t)$ is not orthogonal to t. The idea of sending t to 0, then, is to find some g such that $(g, t) = 0$, but (g, u_{-1}) is still as large as possible. This essentially means making f orthogonal to t. Roughly speaking, one can think of u_{-1} and t as 2 vectors in an inner product space. f is derived from orthogonal projection onto u_{-1}, but since u_{-1} and t are not orthogonal, t has a u_{-1} component. Finding something that will not respond to t, i.e. send it to 0, means finding some new vector v in the subspace orthogonal to t. u_{-1} cannot lie in this subspace, since

it is not orthogonal to t, so the v component of u_{-1} will be reduced. One way to get this is to subtract off the t component of u_{-1}, a canonical orthogonalization. This would be the Gram-Schmidt orthogonalization for t, u_{-1}. In any case, sensitivity is lost. But consider the problem another way. Consider the subspace spanned by u_{-1}, t, and instead of orthogonal projection to a 1-dimensional subspace (the value of a single inner product), look at the orthogonal projection to this 2-dimensional subspace. I.e., try to find the best fit of the form

$$au_{-1} + bt$$

Then a, e.g., can be found by orthogonalizing the basis and renormalizing. I.e. the problem is just that of writing a vector in a non-orthogonal basis. E.g., for the above subspace, one gets

$$a = \frac{(I, \hat{u}_{-1}) - (\hat{u}_{-1}, \hat{t})(I, \hat{t})}{1 - (\hat{u}_{-1}, \hat{t})^2}$$

where $\hat{g}$ denotes g normalized for the appropriate measure and support interval. This is for a single support interval, so it cannot be directly compared with Canny's method, which uses results from 2 different support intervals, without first specifying what measures, support intervals, and subspaces one was interested in.

Using optimization methods analogous to those for his edge operators, Canny also finds optimum detectors for "roofs" and "ridges," and indeed, this could be done for any distribution. Extending these operators to 2 dimensions is somewhat trickier than for the edge. And as the number of operators increases, it becomes harder and harder to make sense of their outputs, since they are not mutually orthogonal. More than 1 operator, e.g., can simultaneously give above threshold output. Furthermore, they may be applied at different scales and at different orientations. Canny calls this problem of understanding all these outputs the "integration" problem, and concedes that it is a sticky one. The problem really stems from the lack of a coherent way to describe the image (locally). Projecting onto some axes that seem interesting is a start, but such a local projection

yields only a few numbers from which it is hard to derive a picture of the qualitative behavior of the image. This is a problem of *dis*integration, or fragmentation. What is required is a coherent way of describing the qualitative structure of the image, in terms of the structures which are of interest.

Linking

Canny uses a fairly effective solution to the "streaking" problem (breaking up of thresholded edges), which he calls "thresholding with hysteresis." While this method does not address the basic problem—understanding the image globally and semantically (in terms of "edges")—it is very workable in the ambient context of ideal edges found by a local linear operator. The technique outputs contours which are the maximal connected contours with some part above a high threshold and all parts above a low threshold. This is equivalent to seeding with strong edges (those above the high threshold) and following the contour at a lower threshold. This is still not quite the same as detecting a weak edge due to its length, *something commonly beyond the ability of edge detectors.* The source of the problem is that one is attempting to find a global object based on local measures. Deciding on a continuation of an edge through a region of poor signal to noise ratio is not a local problem, and it is not clear how to treat it as a signal processing problem. One could try to look for evidence of long straight edges. Canny does this to some extent by using elongated operators. But for much longer edges, this is no longer a local criterion, and special methods are required to deal with the increased combinatorial load (e.g. a directional operator would have to be applied in very many directions). Alternately, a method akin to the Hough transform [Hough 1962] could be used. However, these methods have a bias toward particular shapes of contour; again, since the locality condition is violated, including additional dimensions of shape is very costly. There are semantic problems as well. Consider the case of a long straight edge in a high noise background (or, equivalently, of low contrast). Looking at such data, a human observer generally sees in fact a straight edge in noise. However, it is clear that if

the edge is long enough and the noise is large enough, there must be places along it where an estimator based only on local data will meander slightly to get the best estimate. By the local measure, this will be a better estimate than a long straight edge. That is all one can expect from a criterion which ignores the global shape of the curve.

Empirical results

The results shown appear quite good in that most of the edges of interest are present without a preponderance of "noise" edges. Probably, this is mostly due to using local noise analysis with "hysteresis" thresholding and incorporating the single response criterion (smoothing). While accurate localization is important for applications requiring precision, subjective appraisal cannot take this into account very readily, and the main manifestation of good localization is in localization consistency, i.e. in the estimated location varying smoothly and monotonically with the actual while maintaining minimal scatter. This feature enables, e.g., reliable linking, even if the absolute locations are unreliable. However, here the single response criterion already clears the clutter, so the localization accuracy is probably not very important in this respect.

To the extent that the results are apparently cleaner than other edge detectors, they are very good. However, Canny's results show the same topological problems inherent in all step-matcher filters. These problems are manifested in "wrong" connectivity of contours (i.e., relative to what a human would draw), and occur in places where the image function exhibits a local behavior which is different from the class of functions considered in the design. In Canny's case, the design functions were constants and ideal straight edges, possibly augmented by linear functions. This can be expected to fail in busy places, e.g. corners, resulting in incorrect connectivity, and in fact such behavior is evident in Canny's examples.

Evaluation

This work is a significant contribution to the theory of edge detection by finding extrema of a convolution. Particularly noteworthy are the ideas relating to localization: how to express it mathematically, and how to incorporate a single response criterion. The latter leads directly to smoothing, and thus puts the use of smooth convolution kernels on a firm footing (in computer vision—we are not speaking of statistics in general). The comparisons with other edge operators, e.g. Laplacian, Hueckel, surface fitting are quite useful, as is the discussion of prolate spheroidal wave functions. The ideas have their germ in the work of Marr and Hildreth, but go much further in the way of development, sophistication, and rigor, and are certainly creative on their own.

Without detracting from the quality of the work, the subject matter should be put in some perspective. Each refinement that Canny introduces can be regarded as an enlargement or refinement of some linear feature space, *which is computed pointwise in the image*. By considering sufficiently complicated convolutions, perhaps a great deal can be determined about the image function, although clearly convolution with image-independent kernels cannot yield nonlinear functionals. In any case, for convolutions which are essentially matched filters subject to some constraints, the output of such a filter, at each point, can be regarded as orthogonal projection onto a 1-dimensional subspace of some function space (perhaps after some other linear operation, such as differentiation). The purpose of doing this is to produce a data structure which allows a pointwise decision procedure, e.g. "is there a zero-crossing here?" This works when one assumes the image comes from a very special subspace, e.g. ideal steps. Unfortunately, the type of information required about an image function cannot be adequately compacted to a single number this way. Consequently, further operators must be used to get more information. Thus the feature space is enlarged, but it is very difficult to enlarge it in such a way as to make it easy to solve the decision problem in the feature space. To do this properly, one needs first a theory which says something about equivalence of images. E.g., certain transformations

should not affect the qualitative interpretation of a piece of the image. Another way to put this is that convolving with some number of kernels, and using the information only locally, is the same as projecting some high dimensional space (the pixel values) onto the space of kernels. This is essentially a standard problem of classification. Experience has shown that even with nonlinear classifiers, it is very difficult to duplicate the intuitive distinctions one seeks. The reason for this is that such classification can be incredibly complex unless it incorporates the structure which captures the distinctions. Projection onto an "edge" dimension is an attempt at this, but is not enough. What is required is a more complete understanding of the qualitative shape of the image function, based perhaps on very nonlinear predicates, e.g. equivalence under some class of transformations of the support.

"Feature synthesis" and "feature integration" require goodness of fit. So does incorporation of elongated directional operators. "Non-maximum suppression" can be viewed in the same way. Removal of the response to slow gradients is also a goodness of fit stratagem, in that goodness of fit to a linear function is sent to 0. It seems that the original idea of a single optimal operator has to be modified again and again for different situations. Why is this? There is really nothing wrong with the operator; the problem is with the problem that has been posed. One can find an operator that will respond optimally to a step edge, even for various definitions of optimality and noise process. The difficulty is twofold. First, the definition of step edge is not entirely realistic. There is no reason to expect that a natural edge will be well modelled by a step between regions of constant image intensity (or color). This amounts to a 0th order approximation in the vicinity of the edge. Things may even be worse than just the smooth fluctuations one might imagine, e.g. near the limb of an occluding object, one would expect in principle the derivative normal to the image edge to grow to $\pm\infty$ to a cusp (the intensity would stay finite). Thus the 0th order term would be a bad approximation in any neighborhood of the edge. Deviation from constancy near the step will affect at least the localization from a linear

integral operator with nontrivial support. E.g. $(at+b)u_{-1}(-t)+(ct+d)u_{-1}(t)$, which is 2 ramps of unequal slope separated at the origin by a step, will lead to incorrect localization for Canny's operators, including lateral inhibition. One cannot expect any better, as the feature space includes such a function only in the noise term. Of course, if all such possible functions conform to the hypothesis of Gaussian white noise, the operator will still give the best guess, on the average, but that is probably not what one has in mind.

The second difficulty is that this type of formulation answers the question "If there is an edge there, how can I best determine its parameters?" But, first one really wants to know "is there an edge there?" To answer that, though, besides knowing what "edge" means (not at all a trivial matter), one must be able to estimate how well an edge hypothesis accounts for the data, compared to some other hypothesis. A distribution on one such space (say step edges plus white Gaussian noise) does not translate easily into another distribution on what is essentially the same space (say step edges plus step ramps plus white Gaussian noise). Again this is the problem of selecting the right feature space at the outset.

Unfortunately, if the step edge is not ideal, e.g. if the values either side of the step are not constant, localization based on convolution will be inaccurate. This is easy to see since the extrema of even a linear perturbation $g(x) + ax$ are generically shifted from those of just $g(x)$. This is the manifestation of an intrinsic problem of convolution with smooth kernels: the space of signals is much too large to be adequately classified by a pointwise criterion (such as zero crossing), even for very simple predicates. Restated, it is difficult to find integral approximations to point properties (unless one is willing to integrate against delta functions, of course).

A nonlinear approach

We claim that the problem can be approached through topological methods. The first difficulty is due to the lack of any smoothness in the noise process (as usually formulated,

the noise need not be $\hat{C}^k$ for any k). This would make a frontal assault by topological methods difficult, since the natural settings for such methods are spaces of differentiable functions. However, we can exploit a certain amount of smoothness which is intrinsic to the data. I.e., the data already incorporate some "natural" smoothing, whose effects in any case we would have trouble removing. So we seek to get by with (and exploit) the "minimal" smoothing, without introducing any more confounding convolution. We can consider the data as arising from sampling some smoothing process such as bandlimiting, Gaussian convolution, or even something nonlinear. Since the data space (pixel values) is finite dimensional (though possibly of high dimension), there is a great deal of collapsing in t' e mapping from the infinite dimensional input and model spaces. Therefore, one should seek a regularity condition on the modelling function (i.e., the smooth function we assume gave rise to the data) which will guarantee robustness. Put another way, given an equivalence class of functions yielding the same data, any derived property should be generic. Such a rigorous criterion of robustness is rarely considered; instead, one simply assumes (based on good reasons) that, e.g., the model function is bandlimited. In that case, for a fine enough sample grid, there is a unique (smooth) model function. Since we are not able to provide a rigorous analysis of robustness here, we just consider what happens with a smooth model function, assuming it has been chosen to be generic (or is unique, as with bandlimiting). How can we remove extra extrema, i.e. do smoothing, without blurring? [Koenderink and van Doorn 1979] and [Witkin 1983] have proposed convolving with a 1-parameter family of smoothing operators, and considering how contours of interest (say zero crossings) change as a function of the parameter. One can do this in even more generality, as is done in singularity theory. We can consider generic smooth 1-parameter families of smooth functions, i.e. smooth paths in our model space. In this way, we are led to the theory of generic bifurcations of critical points, since these play the major role in determining the topology of the contours. For real functions in the plane, this theory is very well understood: the only generic changes in the topology of the level set structure are saddle-node bifurcations of critical points, and saddle-connection

nges in the nesting of saddles. Also, the topology of an individual level set can change t passes through a critical point. For functions on the line, things are even simpler: only possibility is max-min bifurcation through an inflection point. One can show t the critical point bifurcations of parametrized Gaussian smoothing are generically dle-nodes as well [Blicher and Omohundro 1984], so that the behavior which occurs Gaussian scale space is described by the usual generic theory. This opens up the pos- ility of doing very non-linear smoothing in a coherent (and simple) fashion. Actually, ce one can now understand the relationships among the extrema, the actual smoothing probably unnecessary. The important thing is that the smoothing must proceed by annihilation of a saddle with an adjacent node (extremum), or the swapping of saddle tings by the saddle connection. This is how *any* smoothing must work. Exactly in at order these things happen for a given function depends on the type of smoothing and perties of the function. Sometimes, this can happen in ways that are not very useful, ce spatial extent and amplitude are interchangeable in a linear (integrating) operator. is probably more useful to know things like the heights, supports, and proximities critical point domains independently. Then certain kinds might be readily smoothed ile others might not, depending on some interpretation heuristic. E.g. even a very ge spike, if of tiny support, could be removed (by excision – i.e. without affecting rby data), or small bumps on a big bump could be regarded as such.

Global Methods

Accumulator arrays

Hough 1962

"Method and Means for recognizing complex patterns" [Duda and Hart 1971, Duda and Hart 1972, Duda and Hart 1973]

The Hough technique offers a solution to the problem of finding global straight lines, or more exactly, finding global sets of nearly collinear feature points. In the present context, "global" means over the entire image, though other workers have used the same idea for subregions.

Basic idea

Consider **L**, the set of all lines in the plane, as a topological space. Duda and Hart use the so-called normal parametrization for **L**, where each line is specified by the pair (θ, ρ), representing the orientation and distance from the origin of the line. This parametrization is borrowed from integral geometry, where it is used in the solution of the Buffon's needle problem. It derives its utility from providing a translation invariant measure for the space, so that probabilities behave in desired ways. ([Santalo 1976] is an excellent source for information about integral geometry, and should be of interest to vision researchers.) Hough, on the other hand, used the slope-intercept parametrization familiar from analytic geometry, but which is fraught with difficulties for this situation. Incidentally **L** is a non-trivial space: for $\rho > 0$, every value $0 \leq \theta < 2\pi$, defines a different line. But when $\rho = 0$, i.e. for lines through the origin, (θ, ρ) defines the same line as $(\theta + \pi, \rho)$. Thus, **L** is homeomorphic to a semi-infinite circular cylinder with the bounded end terminated so that antipodal points on the cross-section circle are identified, which in turn is homeomorphic to a punctured disk with antipodal points on its periphery

ed. This is also the same thing as an infinite Möbius strip, formed by taking a infinite strip and gluing it together with a half-twist.

ısic insight Hough used is this. For each point p in the plane, there is some curve : $\mathbf{L}$ which corresponds to all the lines through p. For each p of interest in the :, accumulate weight for $\gamma(p)$ in $\mathbf{L}$. Then lines in the picture will be places in $\mathbf{L}$ igh accumulated values. (One can think of this as defining a weight accumulation ın $h : \mathbf{L} \to \mathbf{R}$ by $h = \sum \chi_{\gamma(p_i)}$ where $\chi_{\gamma(p_i)}$ is the characteristic function of $\gamma(p_i)$.

da and Hart point out, the method provides a savings because of quantization of e finer the quantization, the less the savings.

ıtion

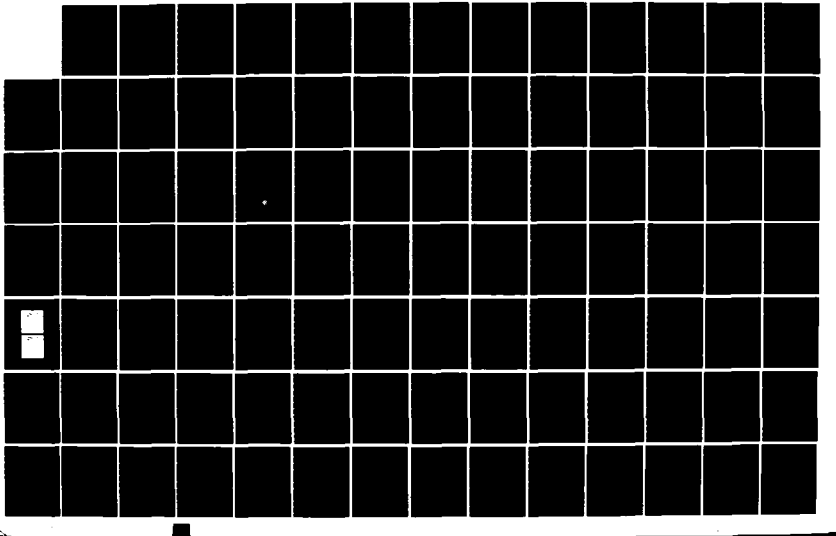
ough method is not adaptable beyond very limited spaces of curves because storage ements grow exponentially with the number of parameters characterizing the fea- i.e. with the dimension of the space of curves.

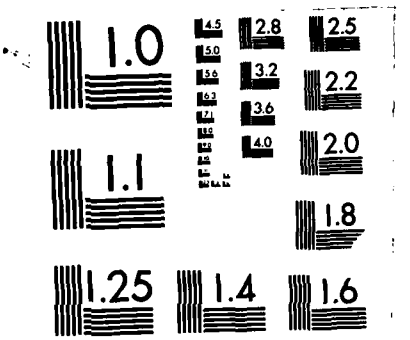
the method is totally global, undesired features can come into play, i.e. the noise high due to many chance contributions throughout the image. However, to combat oblem, one can design localized variations, at the price of requiring a method to *the local results together.*

le generalizations include the detection of curves with more parameters, weighted ulation (based on confidence or significance of the data points), the inclusion of considerations, and localization.

ıccess of the Hough method is very dependent on selecting the initial points of t, i.e. on the local feature operator. On the other hand, a good way of doing this compensate for large parameter spaces.

AD-A155 873 EDGE DETECTION AND GEOMETRIC METHODS IN COMPUTER VISION 2/3
(U) STANFORD UNIV CA DEPT OF COMPUTER SCIENCE
A P BLICHER FEB 85 STAN-CS-85-1041 MDA903-80-C-0102
UNCLASSIFIED F/G 12/1 NL





Ballard and Sklansky 1976

"A ladder-structured decision tree for recognizing tumors in chest radiographs"

The authors are concerned with finding roughly circular regions of approximately 100 pixels in area.

Summary of processing steps

A thresholded gradient picture is first arrived at, using a sequence of processes $\Theta_T \circ \nabla \circ \mathcal{H} \circ \mathcal{L}$, where

$\mathcal{L}$ is the low pass filter operation defined by averaging and then resampling on a coarser grid. (Note that averaging is not strictly low pass, since the filter transfer function is a *sinc*.)

$\mathcal{H}$ is a high pass filtering operation performed in the frequency domain via FFT's, using a filter characteristic attributed to Kruger.

∇ is a digital gradient operator defined in terms of adjacent pixel differences.

Θ_T is a global thresholding operator.

A heuristic search connecting edge pixels, similar to Martelli's technique, is then used to find the lung area, following a Kelly-like "plan."

To locate tumors and nodules, a Hough-like method is used: An accumulator array corresponding to possible circles, indexed by position and radius is incremented by the number of edge pixels with positions and gradient directions consistent with lying on the given circle. An improvement is achieved by using the gradient direction in addition to the magnitude.

Big and small radii are tumor and nodule candidates, resp. The big ones are immediately declared to be tumors, while the candidate nodules are subjected to a 2 stage classifier

which looks at features from a detailed nodule boundary finder. The latter is based on growing all optimal edges of length n in a given region until closure is reached, using a Kelly-like "plan."

Evaluation

The accumulator array method seems to be useful for finding some circle-like boundaries. One must always keep in mind 2 questions for such feature detectors: what does it really find, and what will it miss. These questions are best answered either by mathematical proof or application to numerous examples. Unfortunately, neither of these tests is available in the present paper, though probably one cannot fault the authors for not including more examples, since space limitations may have been imposed. In any case, what is being detected is not regions with roughly circular boundaries, but areas having a sufficiently high count of above threshold gradient values (of the right orientation) lying on a circle. This provides some kind of global understanding of the intensity function, which is commendable, but it is not likely to find sharp edges which do not stay near and tangent to some such circle. However, the main use in the paper being reviewed is to guide a more detailed process of boundary finding, and in that context the question becomes whether the feature being found is indicative of a closed boundary in its vicinity. On the one hand, there is little doubt that a roughly circular boundary of adequate contrast, sufficiently defocussed would cause one or a few such circle detectors to fire, allowing the more detailed process to find the precise boundary. On the other hand, the firing of a circle detector is no guarantee that there must be such a boundary: all that is necessary is that the intensity have a steep enough centripetal gradient over a large enough part of a circle, which might happen if the intensity function has a maximum inside the circle.

Region growing

Brice and Fennema 1970, Fennema and Brice 1970

"Scene Analysis Using Regions"

This is now a classic work in region growing. Its methods are extremely simple, which *a priori* may not be an indictment, but in this case they are based on an overly simplistic image model that no one now believes. The approach was motivated purely by heuristics, rather than any theory, and at this level of processing that turns out to be inadequate.

The basic segmentation operation is to partition the image by pixel intensity value. The authors use boundary predicates which are based on a completely local measure: nearest neighbor intensity differences.

There are 2 merging heuristics:

Phagocyte heuristic

Merge adjacent regions if the "weak" part of their common boundary is a big enough part of one of their total boundaries. "Weak" and "big enough" are relative to global thresholds.

Weakness heuristic

Merge adjacent regions if the "weak" part of their common boundary is a big enough fraction of it (common boundary). Another global threshold is used for "big enough."

Evaluation

The method presented is much too simplistic. E.g., it will clearly fail if smooth shading leads to 1st differences of the same magnitude as an edge. Noise spikes will always end up as regions. The heuristics are too heuristic --they are not based on any analysis or

understanding of real images, beyond a few common-sense notions. Global thresholds are invariably a bad idea: a little observation can persuade one that the same magnitude (of edge parameter, gradient, or whatever) can be significant in one context and meaningless in another.

Kirsch 1971

"Computer Determination of the Constituent Structure of Biological Images"

The author indicates that he is interested in image processing as deriving data structures from image data.

He differentiates between *"well-defined objects"* and *"aggregates,"* which is essentially the difference between smoothly shaded objects with smooth boundaries, and textured "objects" with texture boundaries. He suggests, among other examples, that cells are well-defined objects, while tissues are aggregates.

The goal is to find boundaries for both types of objects, and the approach is via a local contrast function which is based on the use of the convolution masks

$$\begin{matrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{matrix} \qquad \begin{matrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{matrix}$$

and their 90° rotations. The local contrast function C is then defined as the local *max* over all masks of the absolute values from the convolutions. He defines a *blob of heterogeneity* K as (our equivalent definition) a connected region R such that $C|_{intR} \leq K$ and $C|_{\partial R} > K$: basically a low contrast region with a high contrast boundary, with "low," "high" defined by the threshold K.

The data structure he derives is based on the observation that varying the threshold induces a partial order on the regions by inclusion, which is of course a functorial

consequence of the natural ordering on the thresholds. This partial order he represents as a tree, and as a reduced tree showing only when regions coalesce.

Evaluation

The data structure which Kirsch proposes is interesting in that it is essentially the structure of the level sets of the contrast function he uses. As we point out elsewhere, the level set structure of a function captures the topologically invariant information. In this case, however, the preprocessing steps leading to this structure are heuristically based and unfortunately the invariant features are not adequately studied, and the effects of noise on the structure are not taken into consideration. The author cannot really be much faulted for this, as the mathematics involved was not very widely known at the time.

The result is a technique which is better than just intensity thresholding, but suffers many of the same drawbacks. Although he keeps track of what may happen for all threshold values, the thresholds are still *global* thresholds, although one could generalize slightly and use thresholds global only to a region. Now although one might expect boundary contrast to be less variable over some region than simple intensity, it's easy to imagine, e.g., a weak spot in a boundary such that lowering the threshold to include the weak spot introduces enough boundary points to disconnect the region.

Because only the *max* of the directional contrasts is used, important geometric information is discarded in finding the boundaries. This is apt to lead to errors for uniform regions, since noise cannot be rejected on the basis of direction to other boundary points. The effect for textured regions is hard to evaluate; in some cases it may be helpful, but it seems unlikely to work alone.

No justification is given for the values in the convolution mask. For the purpose of detecting a step edge in the presence of Gaussian noise, it is not the most sensitive.

Not enough experimental data is presented to give any feel for performance on real images.

Somerville and Mundy 1976

"One Pass Contouring of Images Through Planar Approximation"

The authors state that they are interested in finding contours of the intensity function, but what they mean by this is finding places of large change of gradient.

Their first goal is to represent the picture data compactly for further processing. This is a good idea, since it is necessary to have a representation of the intensity surface for varying neighborhoods--not just single pixels. An important reason for doing so, which they do not mention, is to synthesize semi-global but accurate information. (By semi-global, we mean regions larger than a single pixel or pair of pixels, yet smaller, usually much smaller, than the entire image.)

The primitive regions to be used for region growing are triangles. These are initially formed by drawing diagonals for each set of 4 points so as to keep similar intensities together. The region growing is then done by a process of raster scan local merging. The merging criterion is as follows.

1) Compute the normal vector of the intensity function on the next triangle. This is not normalized it is actually the 3-dimensional gradient.

2) Compare with the current *average* normal vector for each whole adjacent region.

3) Merge the triangle into the region if the magnitude of the vector error ($|\mathbf{n}_T - \mathbf{n}_R|$) is less than a threshold based on region size:

$$\epsilon_{max}(A) = k_1 e^{-k_2 A} + k_3$$

This can be criticized as follows.

2) Presumably, they do this because they want regions of uniform normal. But it seems more reasonable to compare normals *locally*, leading to locally uniform normal, i.e. regions of slowly changing normal.

3) The adaptive threshold is not well justified. The stated purpose is noise immunity—presumably, values for large regions should be more stable, so there are 2 terms, one for the region noise, one for the triangle noise, though this is not explicitly stated. Since the gradient is a linear operator, one could in fact explicitly solve for the noise characteristics of the expected difference in normals. The region component would be of the form $\sigma = k\sigma_0/\sqrt{A}$, and in fact the standard deviation of error in normals is given by

$$\sqrt{\frac{k^2}{A} + \frac{c^2}{3}},$$

where k^2 is the mean square contribution of each pixel in the region to the expression for the normal, and c^2 is the analogous quantity for a single triangle. In this light, the threshold adopted by the authors is seen to be a linearization and exponential approximation to this function, for a fixed standard deviation of image noise. Furthermore, since the merging is done on a raster scan, the merging predicate will result in different behavior near the tops of regions as compared to the bottoms. Not only that, but this can happen in a discontinous way, when 2 regions suddenly get merged.

The entire process is equivalent to edge detection based on computing a gradient from x and y first differences. However, the edge predicate is adaptive in the sense that the threshold is based on the mean gradient of adjacent regions (in this case, only of regions above, i.e. earlier in scanning). The adaptive part isn't a bad idea, but using an operator with a support of 3 will lead to noise problems, as well as problems with discerning larger scale features.

Experimental results

A single example on a 64 × 48 × 6 picture is given. A reconstruction of the original is presented, based on linear interpolation about the centroid of each region. This result is not impressive. The authors are concerned with data compression and reconstructibility, but from the point of view of image understanding, reconstructibility should not be seen as a measure of performance. The region boundaries displayed do not appear significantly better than other, local, methods. It would be interesting to see the results of a process incorporating the improvements suggested above, viz.

- gradients computed for larger neighborhoods
- thresholds based on picture noise and exact formulas
- merging based on local information. Alternately, one could iterate taking gradients.
- some isotropic merging process (which might result in the requirement for more than 1 pass).

Even so, the gradient idea leads to difficulties if an edge should pass through the operator support – one might get many regions perpendicular to the edge, elongated along the edge, but broken up as the geometry of the edge changes – in other words, poor behavior. A plane is too simple a model for the local intensity surface.

Histogramming

Ohlander 1975

"Analysis of Natural Scenes"

The author does region growing based on analyzing histograms of 9 color image parameters: the 3 raw R, G, B values, as well as the derived parameters of intensity,

hue, saturation and the Y, I, Q parameters used in color signal coding techniques. In addition, values of their gradient as found by a Sobel operator are used, as is the local density of points above threshold in the gradient picture, called the "business matrix." He performs shrinking and expansion on the business matrix to eliminate thin regions (i.e., non-texture edges). The histogram analysis is based on a simple heuristic, and sometimes is done with manual intervention. Regions are found by thresholding.

Evaluation

The technique of thresholding based on features of histograms ignores any geometric relations in the data (a random permutation of the position of pixels doesn't change the histogram). Similarly, it takes no account of the photometric properties of the real world. These problems aside, the use of 9 1-dimensional histograms is still somewhat naïve, since the pixel space is only 3-dimensional. It would be more systematic to use clustering techniques in some 3-dimensional color space (which have an extensive literature) instead of 9 somewhat arbitrary 1-dimensional projections.

This method can be expected to work on images that happen to be amenable to it, i.e. ones where the regions are pretty much homogeneous and separable from others by histogramming. Looking at the technique as a clustering approach, regions can be segmented only if their 3-dimensional pixel color values can be separated by one of 9 families of parallel planes in $\mathbf{R}^3$, the planes perpendicular to the 9 coordinate axes used. This does not even allow for separability by an arbitrary plane in $\mathbf{R}^3$, and the latter is known to already be an overly restrictive condition for most clustering problems.

Shafer 1980

"MOOSE. Users' Manual, Implementation Guide, Evaluation"

Shafer describes a system following Ohlander's technique of image segmentation by the use of multi- spectral histograms. The implementation is essentially automatic, and

reasonably fast (30 seconds on a PDP-10 to segment a 96×128 image, and 20–25 minutes total time with all displays). See the remarks about Ohlander's work regarding the histogramming technique.

The author himself provides some criticism of the technique. The main shortcoming pointed out is referred to as the "majority rule" problem, which occurs when the histogram peak separation process is dominated by large regions. In that case, if a small region happens to be situated in a narrow valley between the large regions (i.e. the large histograms nearly overlap), the small region will be broken in two arbitrarily. This is a consequence of the fact that histogramming ignores geometric relationships. The solution proposed is to first crop the picture so that a small region to be segmented from its surround becomes a large region in the sub-picture. Of course, this amounts to an approximate segmentation. No method is proposed to do this automatically, though the author argues that the cropping idea is robust by showing that including some other objects in the cropped area still allows reasonable performance. This seems to indicate that histogramming works better for very small pictures. A seductive idea (not suggested by the author) is to try arbitrarily subdividing the picture and simply segmenting the smaller pictures. Unfortunately, this will create non- trivial problems in merging regions across subpicture boundaries. In view of the many shortcomings of histogramming and thresholding techniques, it does not seem worthwhile to pursue improvements.

The author also points out the following problems. Many small areas at the boundary of a region are lost since the boundary is sensitive to the threshold. He suggests the solution of merging them after other segmentation is complete. Regions of non-constant intensity cannot be handled, i.e. the technique fails in the presence of any shading. Strangely, he points out that the gradient requires 2 parameters for description, but he does not know how to express this in "one-dimensional features." Presumably, he means he wants to histogram the gradient somehow, but using Ohlander's methods means selecting a single parameter to histogram and threshold. In analogy, e.g. to Ohlander's use of $R + G + B$,

gradient magnitude seems like a reasonable candidate for one such parameter, and it is unclear why the author neglects it.

The eventual goal of this system is for use in an object tracking system. One might hope that even if one couldn't overcome the problems of segmenting a single image, the segmentation would at least be stable from frame to frame. This seems to be a false hope. Thresholding can be thought of as creating boundaries where some level plane intersects the image parameter value function, so that different thresholds correspond to different height contours on a topographic map. At boundaries with small gradient, geometry will change rapidly with threshold value; and at peaks, valleys, and saddles there will be a change in topology as a function of threshold value. If this function has lots of bumps, and if it is changing with time, then there is a serious problem of keeping track of what is going on. The problem becomes one of keeping track of the topological structure of the whole parameter function, particularly its singularity bifurcations, but this cannot be done by simply applying a single threshold, unless the regions created this way are stable over large intervals of time. What can reasonably be expected to be so stable? An object with regions of constant parameter value (shading is tolerable in a system which looks at hue, as long as hue is constant-- an admittedly unlikely situation), moving through light in such a way that the reflectance changes very slowly relative to the motion, against a background having very different spectral characteristics, occurring in an image where everything else also has different spectral characteristics than the object and its background. This appears to be a very limited domain, though there may be useful applications, nevertheless, e.g. in an artificial environment like an assembly line, where these parameters can be controlled.

Optimal linking

Montanari 1970, Montanari 1971

"On the Optimal Detection of Curves in Noisy Pictures"

The author presents a nonserial dynamic programming approach to find optimal 8-connected paths of a fixed length on a grid, and suggests a generalization which permits arbitrary length curves. Examples are displayed with mean square noise = mean square signal ($S/N = 1$) of length 45, with good results, though the examples are not related to real images. "Optimal" is with respect to a figure of merit (FOM); he uses one based on $\sum intensity - \sum curvature$ (he is primarily interested in curves which arise in the character recognition domain).

Evaluation

Using Montanari's method as an edge detector requires developing an appropriate FOM. This is difficult, unless there is a canonical FOM imposed by the problem, since an FOM is not robust in the following sense. Viz., FOM's which are monotonic functions of each other (and as regular as you like) can give different global optima. For edge detection, to the extent that one can estimate the probability that local data were caused by an edge, one can use an FOM based on the relative probability of the curve, so there is promise.

The requirement that the curve be an 8-connected path on a grid is troublesome, since one would prefer smooth curves as solutions. There is no easy way to translate the optimal path to a set of parameters representing a smooth curve, aside from an independent fitting process. Also, it is difficult to take into account any but the most local properties of the curve one is fitting, if for no other reason than the prohibitively large growth of the dimension of the interaction graph for the dynamic programming problem.

Although one is guaranteed an optimum for the FOM, it is not certain that one necessarily wants such an optimum for image understanding applications, at least if the FOM is totally decomposable into spatially local components. The curve one is looking for is one which is the most meaningful in the context of the entire image intensity function (and world knowledge, psychological set, etc.), and this meaning may depend on data away from the curve, which would lead to an intractable interaction graph for a naïve

The condition that the 2 derivatives be nonzero when s is C^r is there for 2 reasons: First, so that the definition reduce to the C^0 case, which is intuitively the meaning of "zero crossing." And secondly, to avoid degenerate cases, e.g. when the locus of zeroes is a submanifold perpendicular to the x-axis in the x, y, θ space, i.e. when the zero locus is tangent to the θ direction. Note that in this case, a zero can still be a regular point of s. Conversely, even if s is C^r, the C^0 condition is weaker, since e.g. it doesn't exclude tangent crossings.

Theorem. $s : \mathbf{R}^2 \times S^1 \to \mathbf{R}$ cannot have an isolated zero crossing in either of the above senses. (By *isolated* we mean there are no other zeroes in the x, θ manifold, for fixed y.) That is, edges cannot be localized simultaneously in x and θ by the zero crossings of a single (θ-parametrized) convolution operator.

Proof.

Case 1: s of class C^r, $r \geq 1$

Since (x, y, θ) is a regular point of s, the implicit function theorem applies and in some neighborhood of (x, y, θ), $s^{-1}(0)$ is a C^r submanifold of dimension 2. The conditions on the partials guarantee that the surface is not normal to any of the x, y, or θ axes, so that for fixed y, there is a curve of (x, θ) values for which $s(x, y, \theta) = 0$, so that the zero cannot be localized in x and θ simultaneously. A more direct way to see this is to observe that what we are seeking is a function s whose zero crossings are the locus of an edge. Regarding the edge as a function $\gamma : \mathbf{R} \to \mathbf{R}^2$, it's obvious that adding orientation leads to a function $\lambda : \mathbf{R} \to \mathbf{R}^2 \times S^1$ defined by $\lambda(t) = (\gamma(t), \theta(t))$, where $\theta(t)$ is the orientation of the edge at $\gamma(t)$. Since the image of λ is 1-dimensional, we cannot hope for it to be the inverse image of a regular value of a map to the reals, since by the implicit function theorem, that must be a 2-dimensional object. But by the same token, if we have instead $s : \mathbf{R}^3 \to \mathbf{R}^2$, then one *can* try to find edges by finding $s^{-1}(0)$. QED Case 1.

The Limitations of Zero-crossings

Definition of zero crossings

It's sometimes handy to have a notation for the function you get by holding the arguments of some other function fixed. We will use the notation $f(\cdot, y)$ for the function that results from fixing the 2nd argument of the function f to be the value y. The dot represents an argument position to be filled. The purpose is to use a notation like $f(x, y)$ while avoiding the confusion of whether it is x, y, or both which are the variable, since each of these cases actually correspond to a differe. t function object. More precisely, suppose that

$$f : X \times Y \to Z$$
$$(x, y) \mapsto z$$

so that $f(x, y) = z$. Now define

$$f(\cdot, y) : X \to Z$$
$$x \mapsto z$$

i.e., $f(\cdot, y)(x) = f(x, y) = z$.

If s is C^0 (continuous), then we will say it has a *zero crossing* at (x, y, θ) if the functions $s(\cdot, y, \theta)$ and $s(x, y, \cdot)$ both have 1-dimensional zero crossings at x and θ, respectively. Colloquially, this means that the x and θ functions have zero crossings. We don't require a zero crossing in the y direction, because it may be the direction of the edge. We will say that $f : \mathbf{R} \to \mathbf{R}$ has a *1-dimensional zero crossing* at x if $f(x) = 0$, x is the only zero in some neighborhood, and f has opposite signs on opposite sides of x in such a neighborhood.

If s is C^r, $r \geq 1$, then we will say that s has a *zero crossing* at (x, y, θ) if $s(x, y, \theta) = 0$ and $D_1 s(x, y, \theta) \neq 0 \neq D_3 s(x, y, \theta)$, where D_i indicates the derivative with respect to the i-th coordinate. Thus (x, y, θ) is a regular point of s, which means that not all of its partials are 0 at that point. This implies the C^0 definition of zero crossing.

Remarks on the definition of zero crossings

The picture we are keeping in mind has the edge oriented along the y-axis. The definitions seem to single out a particular set of coordinates asymmetrically, to keep with this picture. However, the definitions really only require that the x and θ axes *not* be oriented along the edge; equivalently we could have required that *some* set of coordinate axes with these properties exist.

2) For each $y \in U$, each partial $D_j f(x, y)$ (taken with respect to the j-th y-variable) is in $L^1(\mu)$.

3) There exists a function $f_1 \in L^1(\mu)$ such that for all $y \in U$,

$$|D_j f(x, y)| \leq |f_1(x)|.$$

Let

$$\Phi(y) = \int_X f(x, y)\, d\mu(x).$$

Then $D_j\Phi$ exists and we have

$$D_j\Phi(y) = \int_X D_j f(x, y)\, d\mu(x).$$

The lemma permits us to conclude the following

Theorem [Lang 1969]. Let $f \in L^1$ and $\varphi \in C^r$, $r \geq 1$ with compact support. Then $f * \varphi \in C^r$ and $D^p(f * \varphi) = f * D^p\varphi$ for $p \leq r$.

Notice that this means that no matter how badly behaved f may be, $f * \varphi$ is as differentiable as φ. In particular, convolution with a C^∞ function results in a C^∞ function. In our situation, if either the picture or the convolution kernel is differentiable with respect to the parameters (we may interchange the two, allowing the symmetries to act on the picture if it suits us), then our function s, the convolution, is likewise differentiable. If we have a differentiable kernel, then we can take U to be an open set in the parameter space of the kernel, and X to be the picture plane. If instead, we have a differentiable picture, we reverse the roles of U and X. On the other hand it may happen that both the kernel and picture contain discontinuities, e.g. if they have steps. In that case, integration by parts yields the fact that the convolution, i.e. s, is continuous.

Using all the notations, we can include rotations by allowing G to be the rigid motion (Euclidean) group of $\mathbf{R}^2$, and considering the functions defined by

$$S(K,F,x,\theta) = K_\theta * F(x) = (T_{(x,\theta)}(K),F) = (T_{\tau_x \circ \iota \circ \rho_\theta}(K),F) = \int K \circ \rho_\theta^{-1} \circ \iota^{-1} \circ \tau_x^{-1} \cdot F dA$$

We are interested in the function obtained from $K_\theta * F(x)$ by fixing K, F: this is a function $s : \mathbf{R}^2 \times S^1 \to \mathbf{R}$, i.e. a function of x, θ. I.e., we define s by $s(x,\theta) = S(K,F,x,\theta)$. It is the zero crossings of s which we are seeking. Let's underscore the role of the symmetry group G in the definition of s. The construction we used to define s actually defines a map $s : G \to \mathbf{R}$. In fact, for any family of K's defined by some map $M \to \mathcal{F}(\mathbf{R}^2)$, where M is the indexing set for the family, we can define $s : M \to \mathbf{R}$. This way, one can easily add parameters, e.g. to allow different size operators, and this type of analysis is still applicable.

We want to show that for $s : M^3 \to \mathbf{R}$ (where M^3 is a 3-dimensional manifold) the "zero crossings" cannot be an edge locus. To do this, we will have to be more precise about what we mean by "zero crossing," and we will consider separately the cases where s is differentiable, and only continuous. The 2 cases can be analyzed independently; the continuous case subsumes the differentiable case, but since the differentiable case provides better insight, we treat it first.

To begin with, we make some observations about when we can conclude that s is continuous or continuously differentiable. Here is Lemma 2 of Ch. XIV §1 (p. 375) of [Lang 1969].

We take L^1 is the space of all Lebesgue integrable functions, with the norm given by $\|f\| = \int |f| d\mu$. equivalenced by the functions of norm 0. Cf. the definition of L^2 given in the later section on the nonlinear reflection operator.

Lemma. Let X be a measured space with positive measure μ. Let U be an open subset of $\mathbf{R}^n$. Let f be a function on $X \times U$. Assume:

1) For each $y \in U$ the function $x \mapsto f(x,y)$ is in $L^1(\mu)$.

$T_g(K)$ is what you get when you "move K by g;" the right hand side of its definition shows how to calculate the new K. For a translation τ_x, $T_{\tau_x} g(K)(p) = K(p - x)$, which is often baffling for beginning students.

Define the inversion operator, ι, by

$$\begin{aligned} \iota : \mathbf{R}^2 &\to \mathbf{R}^2 \\ x &\mapsto -x \end{aligned}$$

Note that $\iota^{-1} = \iota$, and that in $\mathbf{R}^2$, inversion is the same as rotation by 180°.

Using the notation for inversion and translation, the convolution formula can be rewritten

$$K * F(x) = \int K \circ \iota \circ \tau_x^{-1} \cdot F \, dA$$

where x is now a generic point of $\mathbf{R}^2$ and dA is the area measure. Note $\iota \circ \tau_x^{-1} = \tau_x \circ \iota = (\tau_x \circ \iota)^{-1}$. So, using the T notation,

$$K * F(x) = \int T_{\tau_x \circ \iota}(K) \cdot F \, dA$$

or, abusing the notation somewhat,

$$K * F(x) = \int T_x(K) \cdot F \, dA$$

We can make the notation more compact by using the L^2 inner product $(\cdot, \cdot)$, defined by $(g, h) = \int gh \, dA$:

$$K * F(x) = (T_x(K), F)$$

We can define a rotation operator, ρ, by

$$\begin{aligned} \rho_\theta : \mathbf{R}^2 &\to \mathbf{R}^2 \\ re^{i\varphi} &\mapsto re^{i(\varphi + \theta)} \end{aligned}$$

or, in vector notation

$$K * F(x) = \int_{\xi \in A} K(x - \xi) F(\xi)\, dA$$

We want to use a more abstract notation for this, so that we can generalize it slightly in a transparent way.

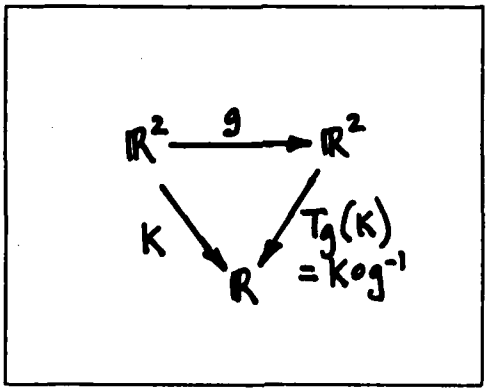


Fig. (β)

Let $g : \mathbf{R}^2 \to \mathbf{R}^2$ be an invertible map, e.g. a rotation or translation of the plane, and let $K : \mathbf{R}^2 \to \mathbf{R}$. To describe "doing" g to K, define the map $T_g : \mathcal{F}(\mathbf{R}^2) \to \mathcal{F}(\mathbf{R}^2)$, where $\mathcal{F}(\mathbf{R}^2)$ is a space of functions each taking $\mathbf{R}^2 \to \mathbf{R}$, by

$$T_g(K) = K \circ g^{-1}$$

Observe that $T_{g \circ h}(K) = K \circ (g \circ h)^{-1} = K \circ h^{-1} \circ g^{-1}$, so the argument transformations "go in reverse order" from the space transformations. Notice also the interesting fact that T_g is a linear map, even if K and g are not. Proof: $T_g(\alpha K + L) = (\alpha K + L) \circ g^{-1}$.

In particular, let G be the translation group of $\mathbf{R}^2$, where $\tau_x \in G$ is defined by

$$\tau_x : \mathbf{R}^2 \to \mathbf{R}^2$$
$$p \mapsto p + x$$

Edge localization in both θ and x

Introduction

One way to localize edges is by finding zero crossings of a convolution operator. This method yields a precise value for, say, the x-coordinate, but to determine the orientation of the edge requires further processing, e.g. using a number of oriented operators (which may disagree as to the x-position) or by observing the locus of zero crossings. An integrated method of extracting the position and orientation would be preferable.

[Binford 1981] proposes localizing edges in direction and orientation simultaneously by convolving a lateral inhibition signal with a directional operator and viewing the results as a set of "stacked planes," one for each orientation of the operator. The estimate for the edge would be based on finding maxima of the gradient of the lateral inhibition signal with respect to position and angle, by seeking zero crossings of the partial derivatives. Since all of the operations prior to finding zeroes would be implemented as convolutions, it is the zero crossings of the resulting convolutions which are sought. As ultimately stated in [Binford 1981], 2 convolutions, corresponding to 2 partial derivatives must be considered. However, it is natural to ask first whether this can be accomplished by finding the zero crossings of a single convolution. In the following, we show that this is impossible, using the inverse function theorem in what is essentially a dimensionality argument. This is why it is necessary for [Binford 1981] to require the use of 2 convolutions.

Some Mathematics of Parametric Convolutions

Let $F : \mathbf{R}^2 \to \mathbf{R}$ be a picture function, and $K : \mathbf{R}^2 \to \mathbf{R}$ a convolution kernel, that we also refer to as a convolution *operator*. The normal definition of the convolution $K * F$ is

$$K * F(x, y) = \int K(x - \xi, y - \eta) F(\xi, \eta)\, d\xi d\eta,$$

this at length in the survey chapter.) We were led to a more general operator, based on symmetry considerations, which turns out to be intrinsically nonlinear. We describe this novel operator, including some of the theory around it, after discussing an idea of Binford's for an operator using a ratio of linear terms, also based on symmetry considerations. The nonlinear operator avoids some of the shortcomings of linear filters.

Finally, we propose a variational technique for combining local edge data into optimal global edges. The key new observation is that the globalization problem, can be put into a form nearly identical to the Lagrangian formulation of mechanics. This allows the global variational problem to be reduced to completely local conditions.

Contributions to Edge Detection

Introduction

In this chapter we present original attacks on some of the problems we discussed in Chapter 2, our edge detection survey. Major problems in edge finding are detection, localization, and globalization; and the most frequent tool is convolution. ***Detection*** consists in determining whether or not an edge is present in a given neighborhood. ***Localization*** is the extraction of the precise position and orientation of the edge. By ***globalization***, we mean finding edge contours of large extent, in contrast to local edge finding, which is concerned only with small neighborhoods. Convolution is commonly used for deriving local information, frequently in a form similar to matched filtering. Greater detail can be found in Chapter 2, the survey of edge detection.

We begin by establishing some background mathematics for studying families of convolution-like operators which are defined by some group, such as rotation. We then use this formulation to prove an original theorem showing that a single family of such operators, parametrized by rotation, is not adequate for simultaneous position and orientation localization by a zero-crossing method. The consequence is that more involved methods, with multiple families, are required for this type of attack.

We present a novel localization operator which uses a least squares fit to find a best local zero-crossing line.

An obstacle in detection is that although matched filters give high response at edges, they also often give above threshold response at uninteresting features. (We have discussed

---

Some mathematical background which is assumed in this chapter, such as functional notation and some results from differential topology, is explained in more detail in the fine print of the chapter Geometric Methods in Vision.

hence only a finite number of final states. One should give some thought as to whether that is an acceptable situation. It could be remedied by altering the relaxation coefficients based on the current state.

be performed by a system of ordinary differential equations. Incidentally, economies can be gained by transforming the state space so as to diagonalize, upper triangularize, or Jordan normal form-alize the relaxation map if it is linear or approximately so. Even if the relaxation is not embeddable in a continuous dynamical system, nearly all the machinery of dynamical system theory is available. For example, if the fixed points are known, theorems are available telling us under what conditions there is convergence near the fixed point, whether the system is stable (i.e. robust with respect to the choice of relaxation parameters), etc.

Evaluation

The experimental results presented are unfortunately not very impressive. But that may well be because the continuous spectrum of labellings generalization has not been made, and because the relaxation (compatibility) coefficients are chosen *ad hoc* without any rigorous consideration of robustness. So the poor experimental results should not be regarded as an indictment of the idea. Rather, it should be developed with greater sophistication. For example, one should consider the effects of noise in a quantitative way. One should try to discover whether there are any global quantities being optimized in the solution. One might consider generalizations to infinite sets of objects, e.g. curves. Thus the local label would be a probability density function for, say, edge orientations and strengths. This leads to an infinite dimensional state space, and although there is a respectable theory of dynamical systems in such spaces, one must confront the computational difficulties. However, since the function factors composing the space are on compact domains, there is a natural decomposition in terms of Fourier series (of the probability densities in the orientation-strength domain), and it is equally natural to truncate these series, so one again obtains a finite dimensional characterization of the state space. One would then want to study the relationship between such a process and, say, variational methods. One can expect that for reasonably regular (finite dimensional) systems, there will be a finite number of fixed points outside of a small neighborhood,

The authors generalize a method first developed in computer vision by [Waltz 1972] for propagating constraints in a graph. Waltz called it "filtering" and used a sequential process; the present authors call it "relaxation" (perhaps due to its similarity to a method used for solving partial differential equations, though it is not derived from it) and do it in an essentially parallel way. One starts with some finite set of objects, some set of interpretations for each object, and a graph where the nodes are the objects and arcs represent mutual constraints between interpretations ("labellings") of the objects. The authors treat 3 types of labelling sets: discrete (finite set of labels) fuzzy (finite set of labels with weights between 0 and ∞), and probabilistic (finite set of labels with weights between 0 and 1). A generalization to a continuous set of labels is not hard to imagine, and would be useful in the applications, for example to represent the orientation of an edge. For the probabilistic case, they readily show that the relaxation process has a fixed point, a necessary condition for convergence. They go on to show that for a class of linear operators with eigenvalues of norm no more that unity, convergence to the unique fixed point is guaranteed. Unfortunately, it's not an interesting case, because the fixed point is independent of initial conditions, i.e. input data. They also present a more interesting nonlinear operator, but are unable to prove that it converges. One can probably invoke one of many variations of the contraction mapping theorem to show convergence for their linear case as well as nonlinear mappings which are contractions in the appropriate sense, thereby expending less effort and achieving greater generality. The important point, however, is that a wide, useful class of such relaxation operators converges. One can even say something about the speed of convergence, based, for example, on the eigenvalues of the relaxation iteration operator. The idea is closely related to dynamical systems, which has interesting implications for neurophysiology and hardware design. If one views the state space as a free vector space on the labels over the field of weights (which we take to be $\mathbf{R}$), then the relaxation is a map of that space to itself. If that map is a diffeomorphism, it may be embeddable as a time-one map of a flow, i.e. it may be the discrete time snapshot of a continuous dynamical system. In that case, the process can

are possible edge elements (pixel adjacencies) and the directed arcs are the allowed edgel successors.

The computation cost varies with S/N, since that is what determines how much searching must be done. This is presented as a positive feature.

He uses a pairwise FOM of edge strength (nearest neighbor difference), but suggests that a larger local operator would improve noise performance.

Evaluation

Summary

The technique is susceptible to the standard problems associated with FOM.
The local operator is still very important.
The same problems as in [Montanari 1970, Montanari 1971] are still present.
The results look reasonable, but no results are presented for real images.
Analysis is required to decide whether the process can be made parallel—
as it stands it is intrinsically sequential.

The technique presented by Martelli in this paper is not usable in its current form. With an appropriate local operator, reasonable FOM, the right discrete variables (i.e. edgel parametrization), it might produce reasonable results. But that says only that global edge finding can be approached as a search problem. Furthermore, it seems likely that parallel search methods would be cheaper (as well as faster) than sequential, in analogy to simultaneous backward and forward searching in classical search problems. An intriguing idea is to use geometric information (i.e. relative direction) of other growing edges to compute the heuristic function (i.e. expansion ordering) for the search problem: edgels would be tried first that led toward something they might mate with.

Rosenfeld, Hummel, Zucker 1975, Zucker, Hummel, Rosenfeld 1977

"Scene Labelling by Relaxation Operations"

implementation (i.e., one without special processing for these other parameters).

The dynamic programming approach is computationally very efficient; generalizations and adaptations of Montanari's method are probably worth pursuing, although it is not a trivial matter to do so.

Martelli 1972, Martelli 1973

"Edge Detection using Heuristic Search Methods"

Following Montanari, Martelli suggests that heuristics should be embedded in a figure of merit (FOM) rather than in code. But it is questionable whether an FOM is enough in the way of heuristics—especially if it is not based on an analysis of real images.

He shows that any dynamic programming problem can be posed as a minimal path in a graph problem, arguing that this is good because the use of heuristics to speed up search in a graph is well-studied. However, the equivalence result is not very deep (each variable expands to a set of nodes, one for each value). The advantage of dynamic programming is that it is far cheaper than graph search, and a better question is usually whether a graph search problem can be cast as a dynamic programming one. One can apply heuristics in the dynamic programming paradigm as well.

The variables x_i of the dynamic programming problem $FOM = f(x_1, \ldots, x_i, \ldots, x_n)$ are the edge elements--discrete valued and thus hard to generalize to continuous θ edgels.

The figure of merit is defined in the form $FOM = \sum c_i(x_i, \ldots, x_{i+k})$ --see the criticism of Montanari that monotonically related c_i's don't lead to the same optimum. In this connection, no discussion of robustness with respect to FOM's is presented.

He derives a search graph for the dynamic programming problem, then uses the A* algorithm to search the graph. The search graph is just a directed graph where the nodes

Case 2: s of class C^{0*}

We restrict attention to the function defined on the $x, 0$ manifold, and show that every zero crossing is an accumulation point of zero crossings.

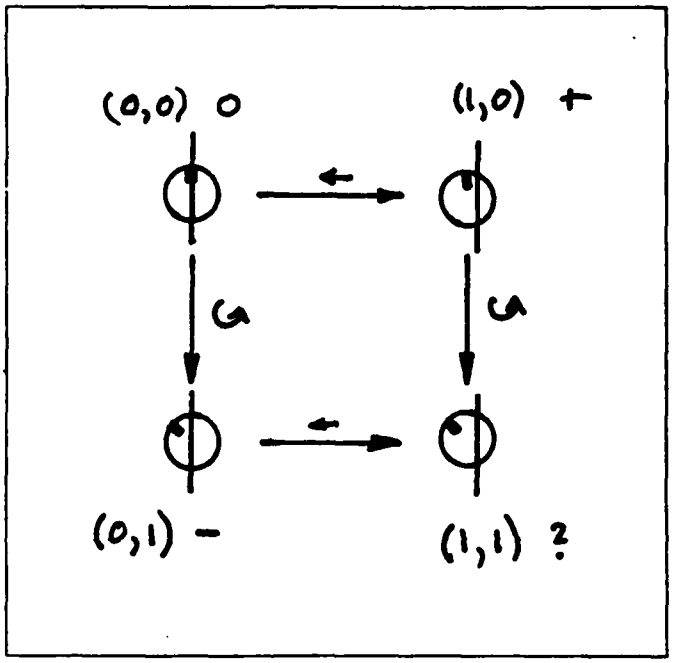


Fig. (proof1)

Look at Fig. (proof1). What it shows, schematically, is an edge operator in the vicinity of an edge, and the result of applying some motions to it. The positions are labelled on an arbitrary scale. The $(0,0)$ position is where the zero crossing is. If one assumes there are no other zeroes in some neighborhood, the indicated operations show that there are 2 ways to get to the same position of the operator with opposite signs for the result, a contradiction.

The notched circle represents the position and orientation of the operator, while the vertical line is a reference value for x and is meant to suggest the edge locus. Starting at the upper left picture, we can get to the upper right picture by translating the operator support to the *left* slightly, which is the meaning of the small arrow over the long arrow.

Since we were at a zero-crossing in x, the value we get by applying the operator at this new position must be nonzero; let us call its polarity +, which we indicate near the coordinates. Instead of translating, we can rotate, and this is schematicized in going from the upper left picture to the lower left. Again, since we start at a zero-crossing in θ, a slight rotation puts us into a nonzero value. Call its polarity −. We can assure that it is not the + of the upper right corner because we have a choice of 2 directions of rotations; by the definition of zero-crossing, one of these will give us + and the other will give −, so we choose the one which gives −. Now we have a contradiction to the assumption that the zero-crossing was isolated, when we observe what happens as we try to get to the lower right corner position. We have assumed that the zero-crossing at the upper left was isolated. Going from the lower left configuration to that of the lower right by a slight translation in the absence of a zero-crossing requires that the polarity of the lower right position be −. On the other hand, doing the same thing by moving from the upper right by a slight rotation to the lower right, *with the assumption* of no zero-crossing, yields a polarity of + for the lower right. The value of the operator applied in the position of the lower right corner can only have one sign, so in fact there must be another zero-crossing somewhere, contrary to assumption. Note that this is still true no matter how tiny the rotations and translations of the operator.

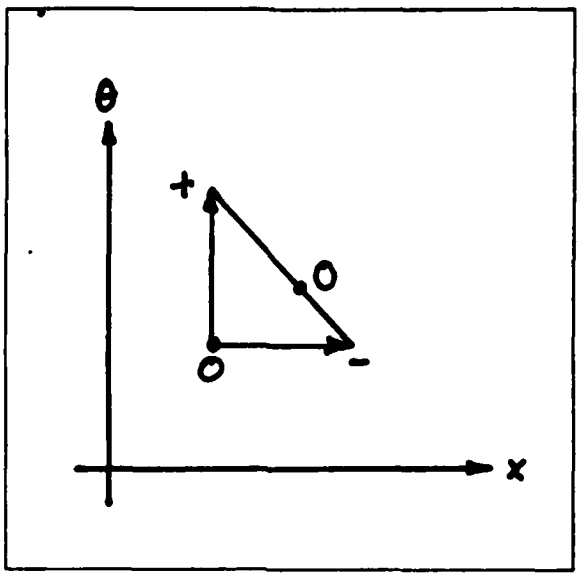


Fig. (proof2)

A more abstract picture of this is Fig. (proof2), which shows a region of the x, θ manifold near the zero crossing. In particular, we can assume without loss of generality that moving up (i.e. rotating) causes s to become + (else flip the picture top for bottom), while moving right (translating) causes s to become − (else flip right for left). The restriction of s to the line joining the 2 end points of these motions must have a zero, by the intermediate value theorem (see, e.g. [Rudin 1964]). **QED**∎

Nonlinear Local Edge Detection

An ideal step function is the sum of an even part—a constant function—and an odd part—a symmetrical step (top of Fig. (latinh)). For edge detection, it is the odd part which is of interest. [Canny 1983], for example, requires that his optimum convolution kernel be an odd function, since the even part cannot contribute to detection of a step. The bottom of Fig. (latinh) shows the (1-dimensional) result of applying lateral inhibition to a step edge, i.e. of convolving a step edge with a zero-sum difference of boxes (middle of Fig. (latinh)).

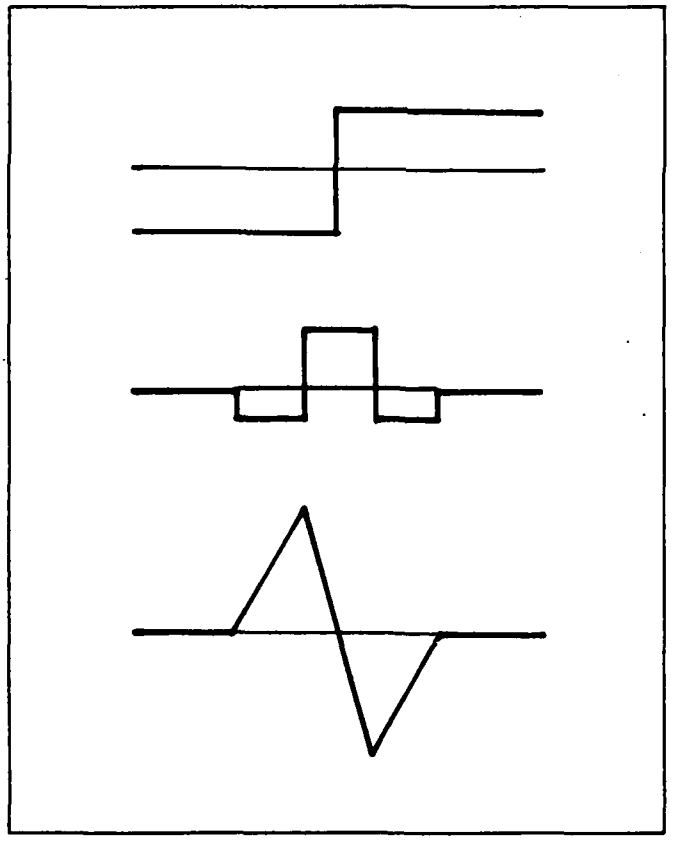


Fig. (latinh)

Since lateral inhibition is an even operator, the result is again an odd function. Also, it is the central zero crossing which marks the edge, while the lateral inhibition has introduced spurious peripheral zero crossings.

A common approach to detecting signals like this is through matched filtering, template matching, or surface fitting, all of which are essentially equivalent linear processes. However, these tend to respond to undesired components while remaining specialized to a particular functional shape. [Binford 1981] proposed using an even-odd characterization for dealing with this problem. We have used a somewhat different characterization of even-odd, which led to the edge detector described below, an intrinsically and nontrivially nonlinear operator. Nonlinearity has the advantage that space and intensity are not equivalent. I.e., a linear operator has no way to tell the difference between a high but localized noise spike and a large moderately positive area. While linearity always has this problem, nonlinearity can avoid it. Also, the even-odd characterization is more general than a matched filter kernel, and thus detects a more general class of functions, so one is not limited to the ideal step. The reflection operator described below will find edges on a checkerboard pattern smaller than its support, something a matched filter cannot do (unless it's a matched filter for that size of checkerboard, of course).

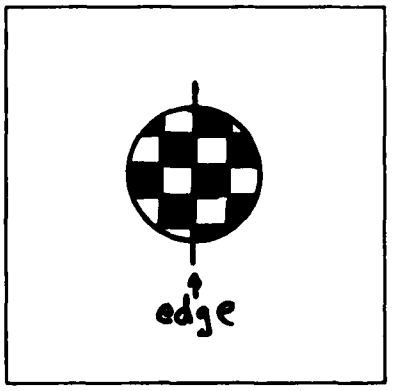


Fig. (checkerboard)

The reflection operator can be thought of as adding up a measure of edgeness along each line perpendicular to the prospective edge, regardless of polarity. This could be done with an operator linear on each such line, if a nonlinear operation such as absolute value or squaring were done before summing the line values. This would result in essentially the same operator, though, once all the nonlinear terms were herded up.

Let f be the laterally inhibited picture function. Then the even and odd parts are defined by

$$f_{\text{even}}(x) = \frac{1}{2}[f(x) + f(-x)]$$
$$f_{\text{odd}}(x) = \frac{1}{2}[f(x) - f(-x)]$$

To make the notation more compact, we can define $\tilde{f}$ by $\tilde{f}(x) = f(-x)$. Then

$$f_{\text{even}} = \frac{1}{2}[f + \tilde{f}]$$
$$f_{\text{odd}} = \frac{1}{2}[f - \tilde{f}]$$

The even-odd operator of [Binford 1981]

[Binford 1981] describes using the even and odd parts as follows. Let

$$R = \int_0^W f(x)\,dx$$
$$L = \int_{-W}^0 f(x)\,dx$$
$$= \int_0^W f(-x)\,dx$$

where we understand that we can consider this a summation by using a discrete measure. Then, in terms of our previous definitions,

$$\begin{aligned} R+L &= \int_0^W [f(x)+f(-x)]\,dx \\ &= 2\int_0^W f_{\text{even}}(x)\,dx \\ R-L &= \int_0^W [f(x)-f(-x)]\,dx \\ &= 2\int_0^W f_{\text{odd}}(x)\,dx \end{aligned}$$

The even-odd measurement is then given by

$$\left|\frac{R+L}{R-L}\right| = \frac{\left|\int_0^W f_{\text{even}}(x)\,dx\right|}{\left|\int_0^W f_{\text{odd}}(x)\,dx\right|}$$

Notice that the only nonlinearity here is in the ratio, and there are no cross-terms in f, since the integration is done over an argument linear in f. Also

$$\frac{R+L}{R-L} = \frac{1+\dfrac{L}{R}}{1-\dfrac{L}{R}},$$

so one is essentially looking at the value of L/R. This computation of even-odd parts has a simple interpretation in terms of convolutions. Define K_+, K_- as in Fig. (R-L).

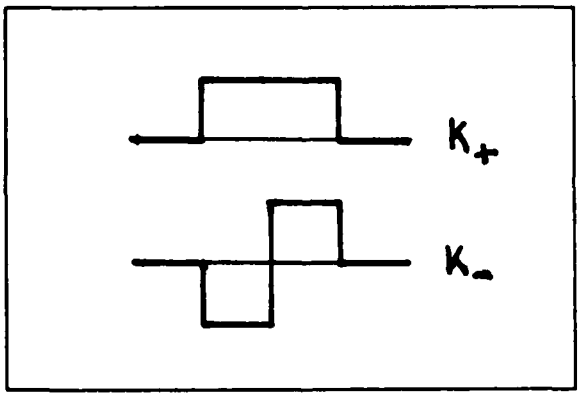


Fig. (R-L)

Then

$$R + L = K_+ * f$$
$$R - L = K_- * f$$

The nonlinear reflection operator

We made a somewhat different interpretation of comparing even and odd parts; we compare the relative sizes of the even and odd parts by considering their norms as elements of a function space. I.e. in our formulation, the detector output is

$$\frac{\|f_{\text{even}}\|}{\|f_{\text{odd}}\|}$$

where $\|\cdot\|$ is the norm coming from an inner product $(\cdot,\cdot)$. For the continuous case, this could be the inner product on L^2, while for the discrete case it could be that on ℓ^2.

We take L^2 as the space of all Lebesgue square integrable real-valued functions, equivalenced by the functions of square integral 0, with the inner product $(f,g) = \int fg\,d\mu$, where μ is Lebesgue measure. The norm $\|f\|$ is then given by $\|f\|^2 = (f,f) = \int f^2\,d\mu$. More generally, one speaks of $L^2(\mu)$, for an appropriate measure μ. If we take μ to be the discrete measure, giving the value 1 at each integer, we get the space ℓ^2, of square-summable sequences.

We then compute

$$\begin{aligned}
\frac{\|f_{\text{even}}\|^2}{\|f_{\text{odd}}\|^2} &= \frac{\|f+\tilde{f}\|^2}{\|f-\tilde{f}\|^2} \\
&= \frac{\|f\|^2 + \|\tilde{f}\|^2 + 2(f,\tilde{f})}{\|f\|^2 + \|\tilde{f}\|^2 - 2(f,\tilde{f})} \\
&= \frac{\|f\|^2 + (f,\tilde{f})}{\|f\|^2 - (f,\tilde{f})} \qquad \text{since } \|f\| = \|\tilde{f}\| \\
&= \frac{1 + \dfrac{(f,\tilde{f})}{\|f\|^2}}{1 - \dfrac{(f,\tilde{f})}{\|f\|^2}}
\end{aligned}$$

Thus it is the relative size of the cross term that is important. The function

$$\frac{1+x}{1-x}$$

is monotonic in x. We are interested only in the relative values of the detector output, and in practice will threshold on its value, so it is sufficient to consider only

$$\frac{(f, \tilde{f})}{\|f\|^2}$$

Note that

$$|(f, \tilde{f})| \leq \|f\|^2$$

and

$$(f, \tilde{f}) = \begin{cases} \|f\|^2 & \text{if } f \text{ is even} \\ -\|f\|^2 & \text{if } f \text{ is odd} \end{cases}$$

The higher dimensional generalization of $\tilde{f}$ we are interested in is reflection across a hyperplane; in 2 dimensions this is reflection across a line. This can be defined by choosing some coordinate system (x, y) and defining $\tilde{f}$ by

$$\tilde{f}(x, y) = f(-x, y)$$

so that we are looking at the odd and even parts along the x-axis. Instead of reflection, we could have generalized instead to inversion, defined by

$$\tilde{f}(x, y) = f(-x, -y)$$

This could be interpreted as looking at the odd and even parts across all lines through the origin at once.

Groups and families of quadratic operators

Using the notation introduced in our section 3.2.2 on parametric convolutions, these operators are of the form

$$\psi_g : \mathcal{F}(\mathbf{R}^n) \to \mathbf{R}$$
$$f \mapsto (f, T_g(f))$$

where g is an isometry of $\mathbf{R}^n$. Since this operation is to be performed at every point of the image, we can parametrize it by a shift as

$$\Psi_g : \mathcal{F}(\mathbf{R}^n) \to \mathcal{F}(\mathbf{R}^n)$$
$$\Psi_g(f)(x) = \big(f, T_x T_g T_{-x}(f)\big)$$

This is roughly the same as our definition of parametrized convolution, except that the fixed convolution kernel K is replaced by the function f itself, giving an operator which is nonlinear in f. For $n = 1$, inversion and reflection are one and the same. For $n = 2$, we have chosen reflection for the group element g. Ψ_g can be thought of as a machine which takes an image as input and gives as output another image, whose value at each point is a measure of the invariance of the input under the symmetry g applied at that point. For example, suppose g is a translation. Then since T_g, T_x, T_{-x} all commute, the value of $\Psi_g(f)$ will not depend on x, and $\Psi_g(f)$ will be the constant function with value $\big(f, T_g(f)\big)$. If we now let g range over all translations, we get a function on the translation group, viz. the autocorrelation function of f. Now let g be reflection across the line ℓ through the origin. Ψ_g takes an input function, and produces an output function. To find the value of the output function at a point x, translate the input function so that x is at the origin, transform the input function by g (i.e. reflect across ℓ), and translate back to x, then take the inner product with the untransformed input function. That's now the value of the output at x. Usually, we are interested in doing this for local support, i.e. the result should only depend on a neighborhood of each point, or be weighted near the point. We can build this into the inner product by using a suitable measure, so that this situation is still described by the same formalism, except it is now more convenient to write

$$\Psi_g : \mathcal{F}(\mathbf{R}^n) \to \mathcal{F}(\mathbf{R}^n)$$
$$\Psi_g(f)(x) = \big(T_{-x}(f), T_g T_{-x}(f)\big)$$

which amounts to taking the inner product at the origin, rather than first translating back to the point of interest. Since translation is an isometry, this does not affect the

value of the unweighted inner product, while for a locally weighted operator, the inner product is just defined once, at the origin.

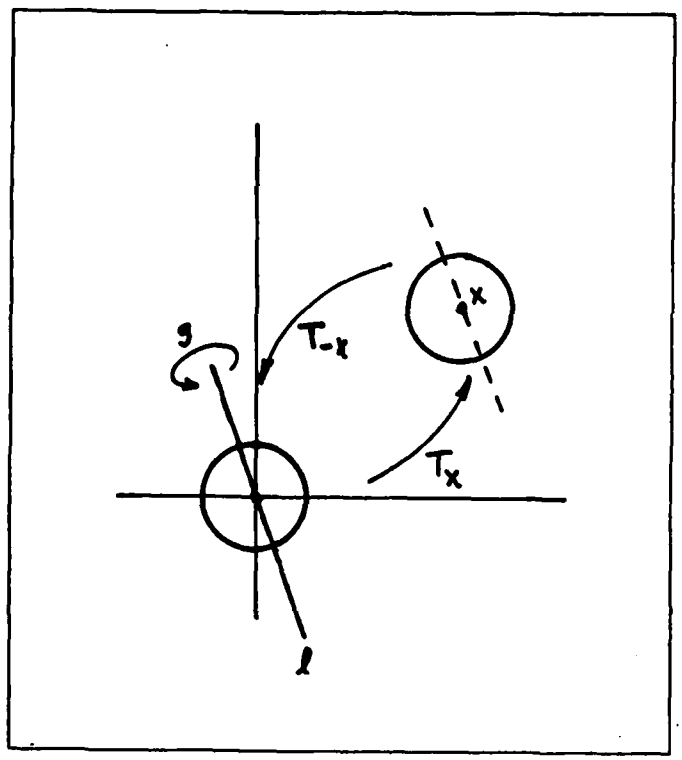


Fig. (Ψ_g)

[Sunday 1978] has shown that Ψ_g is invariant under isometry, in the sense that

$$\Psi_g(T_h f) = T_h \Psi_g(f)$$

for all isometries h, if and only if

$$g \in \text{Center}(O(n))$$

where $O(n)$ is the group of orthogonal transformations of $\mathbf{R}^n$, i.e. $n \times n$ matrices of determinant ± 1, so the condition says that g must commute with all of $O(n)$. Now,

Center$(O(n)) = \{I, -I\}$. I.e. the center consists only of inversion, $-I$, and of course the identity, I. Notice that for 2 dimensions, inversion is the same as a rotation by 180°. Thus, in 2 dimensions, g can only be inversion for Ψ_g to be unaffected by an arbitrary isometry, except of course by being carried along with it. In particular, this means that Ψ_g is not directional:

$$\Psi_g(T_h f)(0) = T_h \Psi_g(f)(0) = \Psi_g(f)(0)$$

If we confine attention only to invariance under rotations, the situation is somewhat different. The rotation group of $\mathbf{R}^n$, $SO(n)$, is the component of $O(n)$ with determinant +1. Since $SO(2)$ is commutative, it is its own center. Thus, non-directional operators in this family could be defined to measure the symmetry of rotation by some arbitrary angle. On the other hand, reflection through a given line is not in Center$(O(2))$ or in the center of $SO(2)$ in $O(2)$ (i.e., all those elements of $O(2)$ which commute with all of $SO(2)$); and the operator Ψ_g it induces is therefore not invariant under isometry or rotation, as should be clear after some reflection. The reflection operator is a directional operator, and must be applied for a family of lines of reflection.

[Sunday 1978] has used essentially the operator above, with g = inversion, for binary pictures as a symmetry detector. For that case, it is interesting that

$$\Psi_{-I}(f)(x) = \frac{f * f}{\|f\|^2}(2x)$$

We note in passing that this could be considered as a 2nd order term in a Taylor expansion in the Fourier domain.

Noise performance

Linear operators mitigate noise by averaging, thus reducing the normalized variance. The nonlinear operators of the class defined above also exploit the correlation properties

of the noise more directly. In the presence of noise, these operators contain a linear and a quadratic noise term. The linear term behaves essentially like that of a linear operator, though it is signal-dependent. It is often assumed that the noise is Gaussian and uncorrelated. In that case, the quadratic noise term vanishes. (I.e., on the assumption that the noise is uncorrelated under the group action involved, its contribution vanishes except on the fixed points of g. In the continuous case, this is a set of measure 0. In the discrete case, this set may not be of measure 0, so care should be taken to avoid including the fixed part, e.g. the line of reflection. It only adds to the noise, and measures nothing of interest. If there has been a preprocessing step, such as lateral inhibition, additional care must be taken, because not all output terms will be uncorrelated.) Thus, while the nonlinearity in the signal term gives a quadratic gain, there is no comparable contribution from the noise. Furthermore, the linear noise term is scaled by the signal, providing additional noise immunity.

Implementation

The nonlinear reflection operator was implemented for a support size of $\sim$ 100 pixels, with uniform weighting. The image was first convolved with a difference of boxes lateral inhibition operator of similar dimensions, with central region of typically 9 pixels. The reflection operator was used only for detection, using a threshold which was set based on a global estimate of noise. The results were qualitatively better than those obtained for various linear detection predicates based on difference of boxes. With tight coding, including automatic compilation of in-line machine code for each operator at run time, thus avoiding any subscript computations later, the cost was essentially the same as for a linear operator of comparable support and nontrivial coefficients.

Planar Fit Edge Location

Applying lateral inhibition [Binford 1981] to a perfect step edge results in a central planar region whose zero crossing line corresponds to the edge locus. We implemented an edge location operator which solves for this zero crossing by finding the parameters of the approximating plane in the appropriate region.

Let L be a lateral inhibition operator, f the input picture, and $L(f)$ the result of lateral inhibition. Define r, s to be the (discrete) coordinate functions in the i, j directions. I.e.,

$$\begin{aligned} r, s: \quad \mathbf{Z}^2 &\rightarrow \mathbf{Z} \\ r: \ (i,j) &\mapsto i \\ s: \ (i,j) &\mapsto j \end{aligned}$$

Then the problem of fitting a plane to $L(f)$ in some neighborhood can be thought of as finding $u, v, w \in \mathbf{R}$ such that ϵ is minimized in the expression

$$L(f) = ur + vs + w + \epsilon$$

Since we are using the ℓ^2 norm with the standard inner product, minimizing ϵ in the least squares sense is the same as minimizing $\epsilon \cdot \epsilon$, which happens if we determine u, v, w by orthogonal projection of L(f) onto the hyperplane in $L(\mathbf{Z}^2)$ spanned by the functions $r, s, \mathbf{1}$, where $\mathbf{1}$ is the constant function. Since we are interested in local fitting, *i.e.*, fitting the central planar region discussed above, the functions $r, s, \mathbf{1}$ must be taken as the restrictions to the region of interest. If this region is symmetrical about the origin, it's easy to see that $r, s, \mathbf{1}$ are all mutually orthogonal, so that the parameters u, v, w are easily found as

$$\begin{aligned} u &= \frac{r \cdot L(f)}{r \cdot r} \\ v &= \frac{s \cdot L(f)}{s \cdot s} \\ w &= \frac{\mathbf{1} \cdot L(f)}{\mathbf{1} \cdot \mathbf{1}} \end{aligned}$$

structure we propose to consider is depicted in the following commutative diagram, Fig. (*). This will require some technical improvements, which we make shortly, but this simpler picture exhibits the main ideas in an uncluttered way.

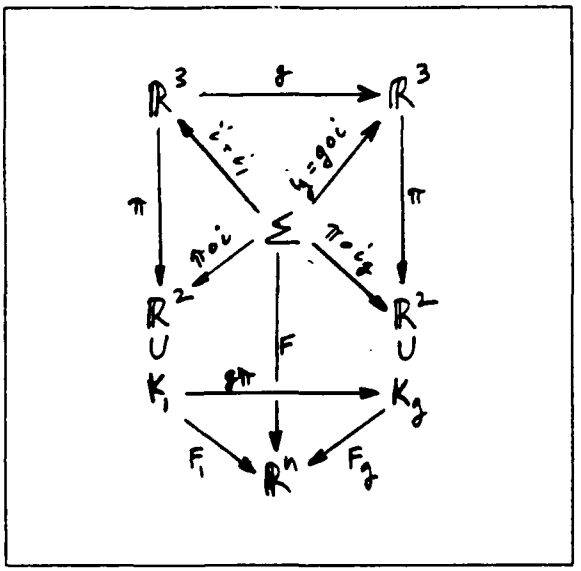


Fig. (*)

An example of functions F_1, F_2 for a neighborhood of a familiar embedding of an object Σ is presented in Fig. (stereo pair). In this case F_1 and F_2 take their values in $\mathbf{R}^1$, which is represented as brightness, and the geometry has been carefully controlled to assure that features will coincide in a particularly simple way (i.e. the images are rectified). This pair is best viewed by holding the page at arm's length, with the pictures side by side, and crossing one's eyes so as to fuse the 2 images into one. This takes some practice.

We will find it easier in our analysis to think of the equivalent situation of a stationary observer in a world which moves. This situation is depicted in Fig. (egocentric example), for a particular choice of object and imaging geometry.

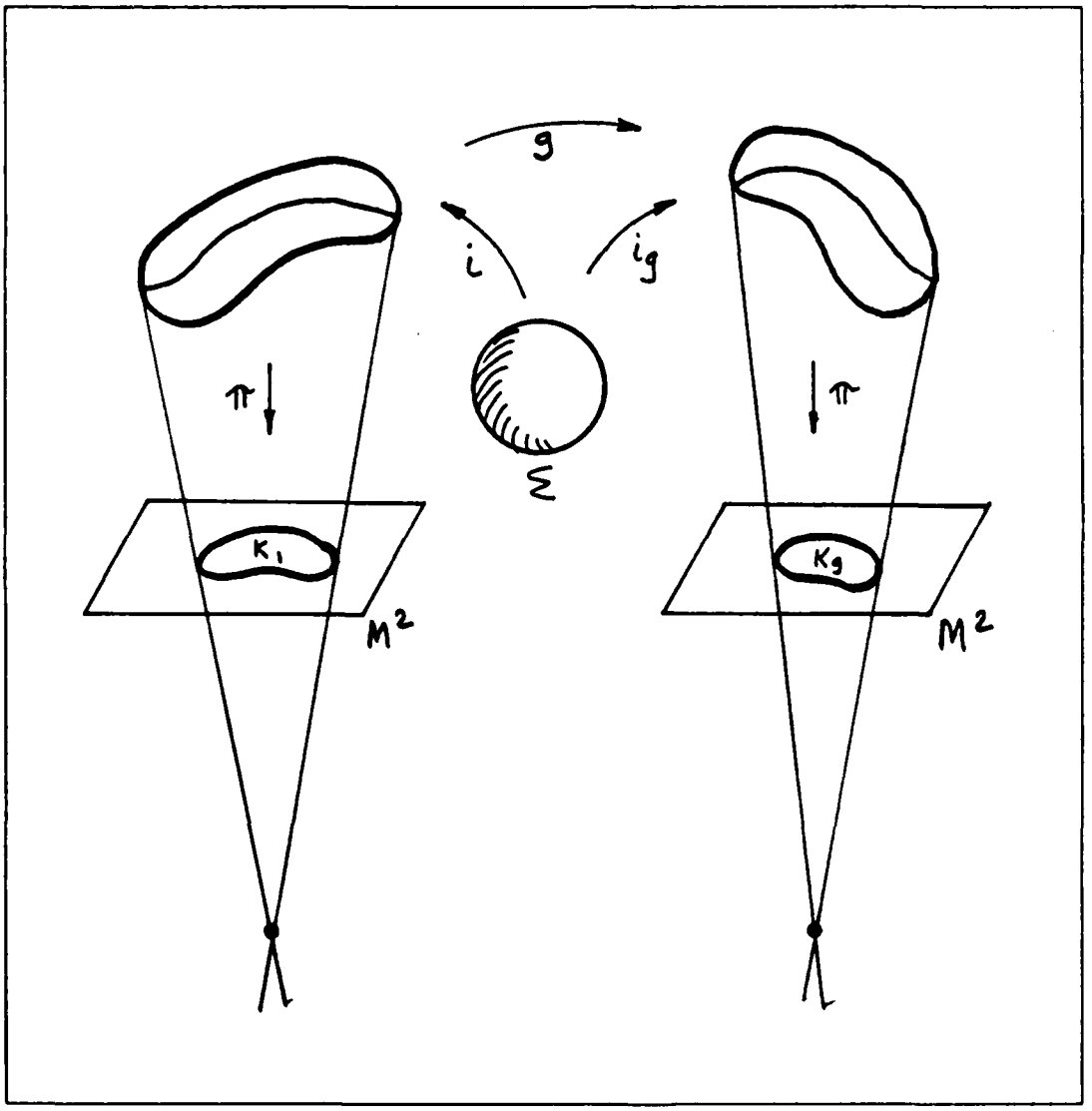


Fig. (egocentric example)

We can formalize the egocentric situation in a mathematical structure which covers a wide range of situations, e.g. different imaging projections. The essence of the mathematical

6) A law which expresses the image intensity as a function of all the other characteristics of the situation.

Fig. (world-o-centric example) is a schematic representation of a possible imaging situation. The name indicates that we are regarding the world as stationary, while the observer moves, which is the usual way of thinking of a stereo imaging situation. The nomenclature is explained in detail later, and is unimportant right now.

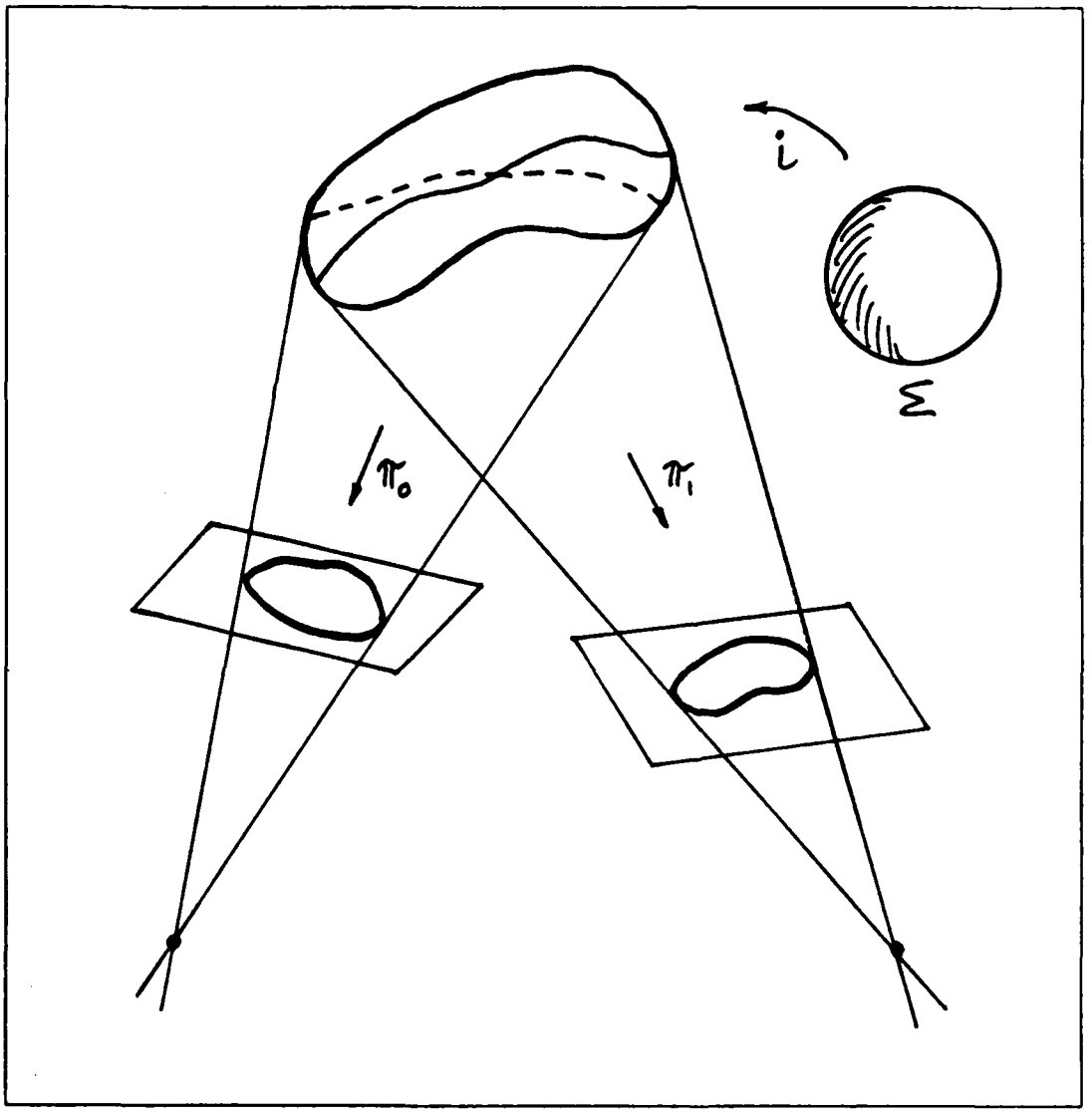


Fig. (world-o-centric example)

The Mathematical Structure

Here is the situation we are confronted with. From a 2-dimensional image or set of images, we want to reconstruct or at least describe the 3-dimensional object that gave rise to our data. Furthermore, we ultimately want to identify objects independent of viewpoint. Now if one can reconstruct the 3-dimensional object, then by brute force one can determine whether 2 data sets (a data set might be a picture, a pair of pictures, a sequence of pictures) correspond to the same 3-dimensional object. However, nature is not profligate in providing us with information, so e.g., one cannot hope to reconstruct the entire object, even in principle; and in practice, accuracy is limited. It would be helpful to know, therefore, something about the likelihood that various data sets may have arisen from the same object, or, more generally, from a single meaningful class of objects. In particular, it would be helpful to know something about how different viewpoints affect the geometry or topology of a data set. So what we have is*:

1) A surface or set of surfaces embedded in $\mathbf{R}^3$.

2) A canonical map from $\mathbf{R}^3$ to $\mathbf{R}^2$ (or possibly S^2), the *perspective projection.*

3) A group of transformations of $\mathbf{R}^3$, viz. the rigid motions of $\mathbf{R}^3$, which correspond isomorphically to the possible ways of viewing an object in $\mathbf{R}^3$.

4) A function defined on the surface, which comprises the intrinsic surface characteristics (e.g. reflectivity).

5) A function defined on $\mathbf{R}^3$, expressing the illumination (which may depend on the embedding and intrinsic surface characteristics as well).

---

*The mathematical notation and some related definitions are reviewed in a fine print section in a few pages. For the moment, it may be helpful to know that $\mathbf{R}^n$ is n-dimensional Euclidean space; and S^n is the n-dimensional sphere, so S^2 is the 2-dimensional sphere--the surface of a solid 3-dimensional ball.

derivative at 6 "typical" points in the picture is just enough. That is for a monochrome picture; in the case of color we show that 3 such points give the same information.

In between the beginning and end of the chapter, there is a middle, in which we apply some differential topology to study invariant structures in pictures. We mainly focus attention on the structure of level sets, the picture loci of each intensity value. These have an invariant tree structure with simple properties given by Morse theory (part of differential topology). We also consider the behavior of the tree in the presence of noise, which again is well understood, and propose the structure as a good starting point for stereo matching. In a later section, we show how the *scale space* paradigm is one way of exploring the structure of the level set tree, and we argue that the invariant structure we propose, including the noise and deformation behavior, is a more complete model for the scale structure of the image, yet it requires no convolutions.

Along the way, we introduce some ideas we need from differential topology, with an eye to explaining their significance in our context of vision. A central idea is *genericity*, a rigorous definition of "typical," which allows us to ignore the problems of special or pathological cases. Without this, our theorems would be impossible, as there would be an endless series of special cases and exceptions to dispose of; instead we can focus on the interesting cases that occur "typically."

territory of differential topology. When we add in the group of rigid motions, we have *differential geometry*.

We apply topological methods to study the correspondence problem of stereo vision, which seeks to find corresponding points in 2 pictures taken from different viewpoints; i.e. matching a point in one picture with the unique point in the other picture that came from the same point on the object (if indeed such a point exists). We begin by assuming that we know nothing of the distortion between the 2 pictures. If we can find this distortion, then we will have solved the correspondence problem. What we find is the novel result (the Two Color Theorem) that this problem is degenerate for monochrome pictures, but uniquely soluble for color pictures (of 2 or more color dimensions). This means that in the monochrome case, the distortion cannot be found without making additional constraints which depend on the properties of the rigid motion group (i.e. the geometry), the projections (including optics), and the possible relation between the viewpoints. On the other hand, for color pictures, we can ignore the geometrical information, or more practically, consider it independently in an overdetermined system. We also extend these results to the situation where the contrast and absolute intensity scale of the pictures may vary, and we consider some of the effects of noise.

At the end of the chapter we return to the geometry which played only a minor role in the proof of the Two Color Theorem, the geometry which we now exploit to analyze the motion problem, the differential analog of the stereo problem. We take the view that our data consists only of pointwise color values in the picture. Since the picture varies with time, we also have pointwise derivative values. Unlike most previous work, we do not assume that we know how any individual points in the image are actually moving (the analog of correspondence), nor do we seek to find that motion as an intermediary to the spatial motion. The question we address, then, is how much of this instantaneous pointwise data does it take to uniquely specify the motion in space. We apply Lie algebra methods to show that knowing the picture function, its gradient, and its first time

Geometric Methods in Vision

Introduction

In 1872, Felix Klein was admitted to the faculty of the University of Erlangen. On this occasion he was required to give an inaugural speech, in which he proposed a characterization of the study of geometry, which had recently seen the introduction of non-Euclidean geometries. His proposal came to be known as the *Erlanger Programm*, and was a unifying influence on geometric thought for the next 50 years or so. The essence of what he suggested is that a geometry should be viewed as the study of the invariants of the action of some group. In Euclidean geometry, for example, one studies invariants of the group of rigid motions of the plane. One can view various geometrical studies this way, e.g. special relativity considers the invariants of the Lorentz group, while topology studies those of groups of homeomorphisms. In the same spirit, the task of computer vision can be viewed as finding invariants of picture functions under the rigid motion group of 3-dimensional Euclidean space.

As an object moves in space, or as we change our viewpoint, the projection of the object's points to the picture undergoes a deformation which depends on the shape of the object, the motion, and the projection. Carried along with this deformation is the picture function, given by the color value at each point, which is a result of intrinsic properties of the solid object, but which can depend on lighting conditions in addition to the deformation of the projection.

Our first goal in this chapter is to make precise what are all these functions, objects, projections, and motions, and what are their relationships; in other words, to describe this structure in the language of modern abstract mathematics, giving us something to attack with rigorous tools. The structure we find, of manifolds and maps, is the natural

excursions of the trajectory, as might happen in trying to maximize the integral of a positive quantity (such as $p(h(x) \mid E(x,\dot{x}))$). In maximizing a positive quantity, lengthening the curve always increases the integral, so e.g. extending the curve always improves things, and one can have the pathology of improving a curve by taking out a tiny piece and replacing it by some wild excursion which accumulates more of the positive goodness. Minimizing a positive quantity (or maximizing a negative one) avoids this, since there is a shortest path, i.e. one cannot keep minimizing by always shortening the path.

In summary, the outcome is that the Lagrangian picture allows us to reduce the extremal problem to a local one. Since we can estimate $\partial L/\partial \dot{x}\,(x,\dot{x})$ and $\partial L/\partial x\,(x,\dot{x})$, we can find, numerically at least, the trajectories that solve the Euler-Lagrange equations, and this is based on local information. The key features making this possible are the existence of the Lagrangian function defined on the space of $(x,\dot{x})$, and the constraint that the only trajectories of interest in that space are those where $dx/dt = \dot{x}$.

Since exponentiation is monotonic, extremizing the exponential is equivalent to extremizing the exponent. This leads to a simple way to extend this to the continuum, by generalizing the sum to an integral (as could be done for any product). The condition then becomes one of maximizing

$$\int \log p\Big(h(x(t)) \mid E(x(t), \dot{x}(t))\Big)\, dt$$

which is a negative quantity, or, perhaps more intuitively, of minimizing

$$-\int \log p\Big(h(x(t)) \mid E(x(t), \dot{x}(t))\Big)\, dt$$

I.e., we can choose $-\log p\big(h(x) \mid E(x, \dot{x})\big)$ as the Langrangian $L(x, \dot{x})$.

Integrating the Euler-Lagrange equations requires an initial condition (or possibly a boundary condition). Since the space in which the equations are set is the $(x, \dot{x})$ space, the initial condition must specify ***both*** x ***and*** $\dot{x}$. In general, different initial values of $\dot{x}$ will give different trajectories. That is the price one pays for getting a completely local problem. However, this can be readily dealt with by separately maximizing over directions of $\dot{x}$, or choosing initial $\dot{x}$ at points of high confidence (seeding). Alternately, the phase portrait associated with the trajectories can be thought of as a "primal sketch" of the potential global edge structure of the image. This structure directly represents simultaneous multiple, even conflicting, interpretations. E.g. there may be more than one value of $\dot{x}$ at some x or in some neighborhood, which gives a tenable edge locus. The orbits, i.e. global edges, have a measure assigned to them by the Lagrangian integral, so there is a ready way to rank and prune multiple interpretations.

As long as "edgeness" does not have a canonical definition, we can't avoid a heuristic aspect to the choice of Lagrangian. The particular extremal problem we have suggested, however, has the nice property that the integral value cannot be improved by arbitrary

the magic of the calculus of variations, this can be reduced to a purely *local* condition on the trajectories, given by the Euler-Lagrange equations: $d/dt\,(\partial L/\partial \dot{x}_i) - \partial L/x_i = 0$. I.e., the solutions to the variational problem can be found by solving the system of equations

$$\begin{aligned} \frac{dx}{dt} &= \dot{x} \\ \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right) - \frac{\partial L}{\partial x_i} &= 0 \end{aligned} \qquad \text{(Euler-Lagrange)}$$

Thus for our situation, all we must do is define an appropriate Lagrangian function $L(x, \dot{x})$. Of course, this will be related to the local "edgeness" function. Typically, an "edgeness" function is the result of applying an operation which measures the degree to which the image *locally* resembles an edge. E.g., one might convolve with a family of optimal filters, such as oriented smoothed steps; the output would be an "edgeness" function depending on position and orientation.

We describe one candidate for such a Lagrangian function. Suppose that at each point x of the image we have computed some information, perhaps by convolving with some set of operators; call this information $h(x)$. Define $E(x, \dot{x})$ to be the event that there is a local edge of magnitude and direction $\dot{x}$ at the point x. Then with some assumptions about the noise process we can estimate $p\big(h(x) \mid E(x, \dot{x})\big)$, the probability density that $h(x)$ arose as a "consequence" of $E(x, \dot{x})$. The function $p\big(\cdot \mid E(x, \dot{x})\big)$ is a probability density on the space of data $h(x)$, for each $E(x, \dot{x})$. (Note that we have no *a priori* estimates for $p\big(E(x, \dot{x})\big)$, and that the entire event space need not be $\cup_{x,\dot{x}} E(x, \dot{x})$.) Then we can argue that for a set of points along a contour, we want to maximize the resulting joint probability density for all the points. Assuming independence, this becomes

$$\prod_t p\Big(h\big(x(t)\big) \mid E\big(x(t), \dot{x}(t)\big)\Big)$$

for integer t, i.e. a finite set of points. This can be conveniently rewritten as

$$\exp\left(\sum_t \log p\Big(h\big(x(t)\big) \mid E\big(x(t), \dot{x}(t)\big)\Big)\right)$$

A Variational Principle for Edge Linking

The field of edge detection has seen no particularly successful consideration of global shape (though see [Marimont 1984] for some recent work in that direction). One can try to find global edge contours either by solving for global information directly (e.g finding a level set of some function), or by piecing together data from simple local operators, as [Montanari 1970, Montanari 1971, Martelli 1972, Martelli 1973] did. Here we offer a variational approach for the latter kind of contour finding.

The essence is the observation that there is a formal similarity between optimal edge linking and Lagrangian mechanics.

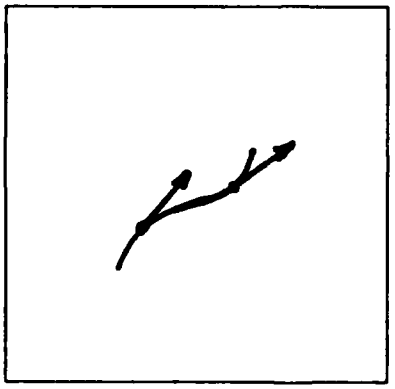


Fig. (path)

Consider a trajectory $\gamma : I \to \mathbf{R}^2$ in the image, which we think of as an edge locus. We can represent a *local* ideal step edge at each point $\gamma(t)$ of $\gamma(I)$ as a vector whose direction and magnitude represent those of the edge. (By magnitude of edge, we mean the size of the step.) This establishes a correspondence between trajectories in the plane and edge loci.

In the Lagrangian picture of mechanics, the state space (phase space) is a $2n$-dimensional space of $2n$-tuples $(x_1, \ldots, x_n, \dot{x}_1, \ldots, \dot{x}_n)$, and the trajectories through this space must extremize the integral $\int L(x, \dot{x})\, dt$ and satisfy the constraint that $\dot{x} = dx/dt$. Through

to the true and precise location of the real edge giving rise to the data. Local detection of edges is not an end in itself, but only the first step in the process of contour finding. The process of assembling the local edges into contours will confront ambiguities where it is not clear which, if any, contour a local edge belongs to. The coarser the resolution of the local edge parameters (e.g. position, orientation) the more frequently ambiguities will arise. As long as we compute the local edge parameters as nonsingular smooth functions of the true edge parameters, then the computed values will be samples of a continuous function, and subpixel resolution will serve to decrease ambiguity.

−1	0	1	−4	−4	−4	1	1	1
−1	0	1	−3	−3	−3	1	1	1
−1	0	1	−2	−2	−2	1	1	1
−1	0	1	−1	−1	−1	1	1	1
−1	0	1	0	0	0	1	1	1
−1	0	1	1	1	1	1	1	1
−1	0	1	2	2	2	1	1	1
−1	0	1	3	3	3	1	1	1
−1	0	1	4	4	4	1	1	1
	r			s			1	

to find the parameters for a vertically elongated region (with analogous operators for other directions). Square operators could also be used.

These convolutions yield at every point $p \in \mathbb{Z}^2$ three parameters $u(p), v(p), w(p)$, determining a plane given by $z = ux + vy + w$ which is the best fit to the data $L(f)$ in the translated support of the convolution operators. The position and orientation of the edge, i.e., the parameters of $u_{-1}(ax + by + c)$, are given by finding the zero crossing line of the fitted plane, i.e. by solving $0 = ux + vy + w$.

This operator gave qualitatively good results for location and direction of edges in numerous real pictures.

Subpixel localization

The zero crossing parameters found by the above method give an edge locus to subpixel precision. For an ideal edge, with sufficiently low noise, this is an accurate estimate. Real edges are not ideal, and it would be quite fortuitous if the nonideality occurred in just such a way as to make the subpixel estimate accurate. Nevertheless, making such an approximation for subpixel location is useful, even without knowing that it corresponds

Of course, this is only good for a region centered at the origin. For a region with arbitrary center, we can apply the same technique modulo a translation to the origin. Equivalently, since we are talking about a family of regions congruent under translation, we can consider u, v, w to be functions on $\mathbf{Z}^2$ expressing the parameters of the plane fit in local coordinates centered at their argument. Then we have

$$u = \frac{r * I(f)}{r \cdot r}$$
$$v = \frac{s * I(f)}{s \cdot s}$$
$$w = \frac{\mathbf{1} * I(f)}{\mathbf{1} \cdot \mathbf{1}}$$

This permits us to implement the least squares fit as convolutions with the functions $r, s, \mathbf{1}$.

For example, using the lateral inhibition kernel

$$\begin{array}{ccccccccc}
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & 8 & 8 & 8 & -1 & -1 & -1 \\
-1 & -1 & -1 & 8 & 8 & 8 & -1 & -1 & -1 \\
-1 & -1 & -1 & 8 & 8 & 8 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1
\end{array}$$

we can use $r, s, \mathbf{1}$ masks

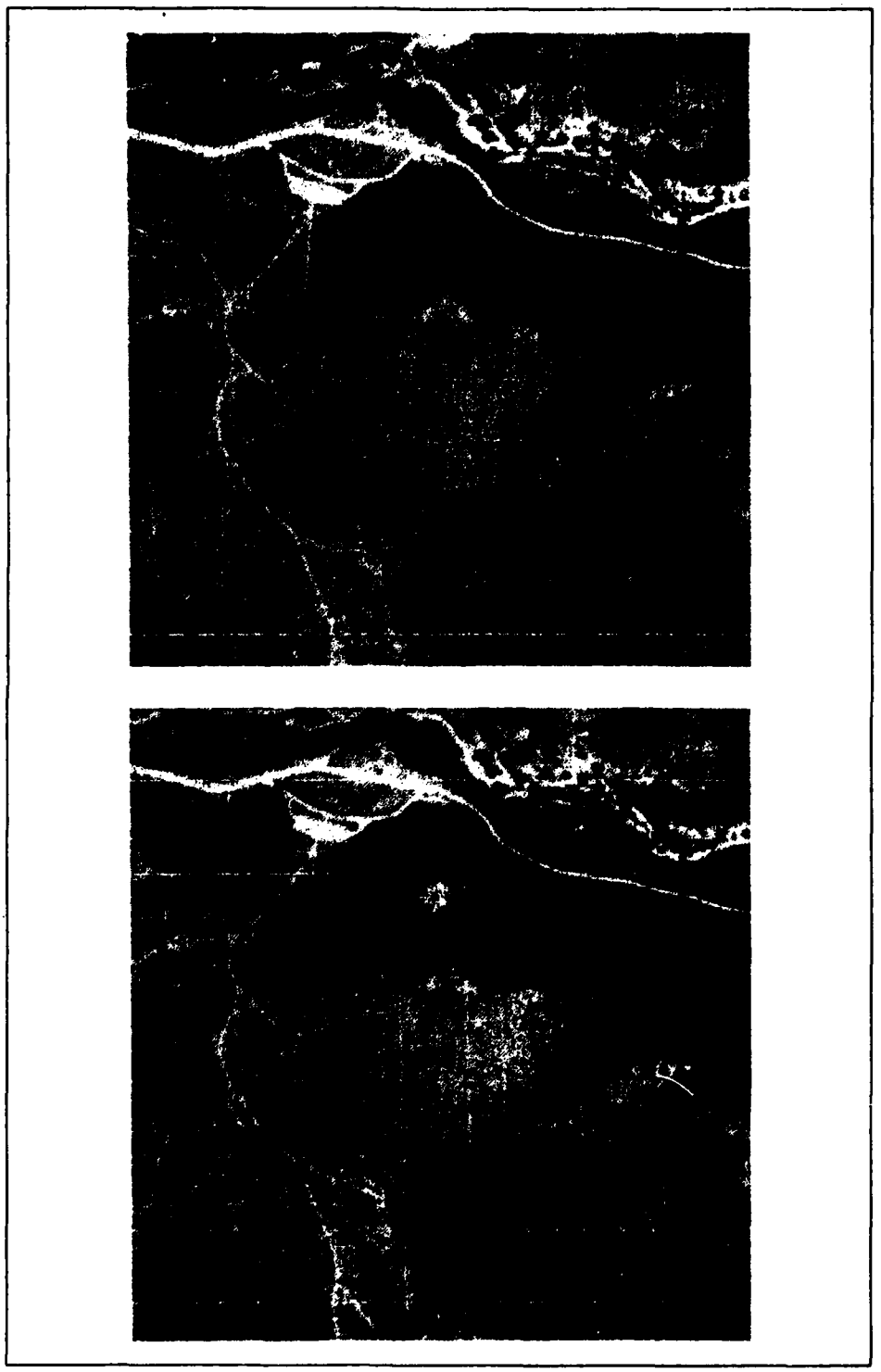


Fig. (stereo pair)

Although the diagram Fig. *(*) may at first appear rather formidable to the non-mathematician, it is actually fairly simple. Nevertheless, here is a detailed explanation.

The diagram represents the relationships among various maps between various spaces. To be exact, the symbols at the nodes represent spaces, the arrows represent maps (i.e. functions) from one space to another, and the symbols along the arrows name the maps. Sometimes a map can also be thought of as a point in some other space, but that is not represented in the diagram. Saying the diagram is *commutative* means that any path along arrows (concatenated by composition) joining two spaces gives the same result. E.g. the following diagram

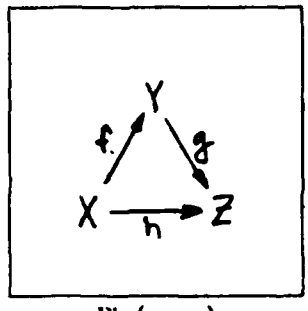


Fig.(comm)

is commutative iff $h = g \circ f$. Diagrams which are not commutative are generally confusing.

The problem of defining or representing the general surface in spaces of dim ≥ 3 is not trivial. That is because the surface may have a strange configuration, e.g. it may close on itself like the sphere or torus, or it may wind around itself.

The simplest examples of surfaces are given by equations of the form $z = f(x, y)$, i.e. as a map $f : \mathbf{R}^2 \to \mathbf{R}$ which we can interpret as assigning to each point on the plane a height above the plane.

There is at present a divergence between the mathematical literature on the one hand, and the engineering and scientific literature on the other, with respect to the notation used to represent functions. This is a divergence which has developed during the past several decades, primarily because of the mathematicians' realizations of the necessity of making explicit the existence of a function (or map) *as an object in its own right*, as well as the requirement of avoiding various ambiguities which otherwise arise. In the engineering literature one often sees references, e.g., to

"a function $z = f(x, y)$"

For many purposes, it is clear enough what this means. However, to be precise and avoid confusions we will adhere to the following notations.

$$f : X \to Y$$

will mean that f is a function which maps points in the space X to points in the space Y. (By function, we mean a single-valued function, or an assignment rule.) Additionally, the notation

$$f : x \mapsto y$$

or

$$f : X \to Y$$
$$x \mapsto y$$

will mean that f takes the point $x \in X$ to the point $y \in Y$, which we will also write as

$$y = f(x)$$

Note the difference between, e.g. x and X, and especially the different meanings of the 2 types of arrows. In this notation, rather than saying

"the function $z = f(x, y)$"

we will say

$$f : \mathbf{R}^2 \to \mathbf{R}$$

$$(x, y) \mapsto z$$

(Here $\mathbf{R}^n$ is n-dimensional Euclidean space). Here the function is f, which is a map from $\mathbf{R}^2$ to $\mathbf{R}$. $f(x, y)$ is the *value* of the function f at the point $(x, y) \in \mathbf{R}^2$. Notice that we might have said, e.g.

$$f : S^2 \to S^1$$

$$(x, y, z) \mapsto \theta$$

(S^n is the n-dimensional sphere given by $1 = \sum_{i=1}^{n+1} {x_i}^2$, where the x_i are coordinates in $\mathbf{R}^{n+1}$. S^2 is the 2-sphere (homeomorphic to the surface of a ball) and S^1 is the circle.)

We can think of this as a surface by considering the points of the surface as given by the graph of f, i.e. by $\{(x, y, z) \in \mathbf{R}^3 \mid z = f(x, y)\}$. We can describe the surface as a function $\hat{f} : \mathbf{R}^2 \to \mathbf{R}^3$ given by the formula $\hat{f}(x, y) = (x, y, f(x, y))$. Unfortunately, most surfaces, e.g. the sphere, cannot be described this way. Most importantly, no matter where we place the plane, there are usually either 2 points or 0 points of the sphere above any point on the plane. One can remedy this by defining the surface as $\{(x, y, z) \mid g(x, y, z) = 0\}$ for an appropriate function g. E.g., for a sphere of radius r one would take $g(x, y, z) = x^2 + y^2 + z^2 - r^2$. It turns out that one can essentially get all surfaces this way, but there are unpleasant side effects which cause us to avoid this definition. To guarantee a meaningful concept of dimension, we would have to impose extra conditions. Besides, finding the set of points that make up the surface is hard. Instead, we define a surface by observing that a little piece of it is very much like a little piece of the plane. We define a *patch* of the surface $\Sigma \subset \mathbf{R}^3$ to be a smooth 1-1 map $\varphi : U^2 \to \mathbf{R}^3$, where $U^2 \subset \mathbf{R}^2$ is a *neighborhood* (i.e. an open set) in $\mathbf{R}^2$, such that $\varphi^{-1}|_{\varphi(U^2)}$ is also smooth:

Fig. (patch)

A surface is then defined to be a collection of such patches such that for any 2 patches (φ, U), (ψ, V), $\varphi^{-1}(\varphi(U) \cap \psi(V))$ is an open set in $\mathbf{R}^2$. This guarantees that our object is uniformly 2-dimensional and doesn't have self-intersections. Such a patch is often also called a *chart* in analogy to the charts of the earth's sphere used by mariners. Similarly, a compatible collection of such charts, covering some object, is known as an *atlas*.

By *smooth* we mean continuously differentiable some number of times. In particular, we use the notation C^0 to represent the class of 0-times continuously differentiable functions, i.e. continuous functions; the notation C^k for k-times continuously differentiable functions; C^∞ for infinitely differentiable functions; and C^ω for analytic functions, i.e. infinitely differentiable functions representable by a Taylor series. Usually *smooth* means C^∞, but sometimes it can mean C^k for some (finite) k. Usually it is immaterial, but if it matters it will be stated explicitly.

A *homeomorphism* is a 1-1 continuous map with a continuous inverse. A C^r *diffeomorphism* is a C^r homeomorphism with a C^r inverse. Often we will not specify the degree of smoothness of a diffeomorphism, as it may not be important or it may be clear from context. Nearly everything we consider can be thought of as C^∞, and we will state when this is not so.

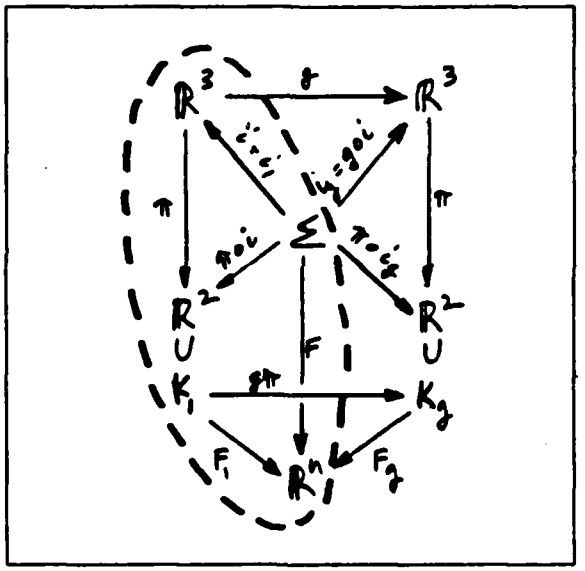
Fig. (*) is meant to capture some of the basic features of imaging geometry. The object surface we are looking at, Σ, sits in 3-dimensional space. That 3-dimensional space has a standard projection, e.g. perspective projection, to the image plane, a 2-dimensional space. Our data, i.e. the picture, is not the projection of the object in the image plane, but rather a color or brightness function defined on that projection. Meanwhile, we can change our viewpoint, or the object can move. We view this "egocentrically," as we explain shortly, and take this as a motion of the whole world while we stay put (we are not considering relative motions of various objects); this motion is g, and the object is carried along with it, rigidly fixed in 3-dimensional space.

Now we must consider what happens to the picture function. The complete physics of the situation is that the observed picture changes as a function of the surface orientation, the lighting direction, and the observer position, as embodied in the image irradiance equation, in addition to undergoing geometric distortion. We have lumped the *photometric* considerations together into a constant effect on the observed image irradiance, to keep things as simple as possible for an initial analysis. They could readily be included by using a sphere bundle over the surface, for example, to account for the relative positions of observer, surface, and light. The simplifying assumption we have made, then, is that the photometric effects of change in viewpoint are negligible in comparison to the geometric ones. This is frequently a reasonable assumption, as is evidenced by the fact that we do not often experience the retinal rivalry which occurs when the assumption is violated. Of course, the other extreme occurs with specular reflection, when the photometric effects are dominant. The consequence of this assumption is that observed data values are carried along with the object. All matching systems we are aware of to date are also predicated on this assumption in that they deal only with features rigidly attached to objects. Some can mitigate some photometric effects by using features such as "edges," but edges of specular reflections are still disastrous.

We assume, therefore, that the picture function we detect is rigidly fixed to the object

surface. This can be thought of as associating picture point values with points on the object surface, although these values are really derived from intrinsic surface characteristics and the image irradiance equation. This fixed association is specified by the function F. Then the distortion g_π between images tells us how the 2 pictures F_1, F_g are related. We want to study the problem of finding the distortion g_π and the motion g just from the data F_1, F_g.

We now lay this out in more detail. First let's consider just a part of Fig. (*), shown in Fig($\frac{*}{2}$).



Fig($\frac{*}{2}$)

Roughly speaking, here is what we are depicting. The surface Σ is embedded in $\mathbf{R}^3$ via i. π is the imaging projection from $\mathbf{R}^3$ to $\mathbf{R}^2$, the image plane. F_1 is the observed image intensity on some closed set K_1 of the image plane, and F is the intrinsic surface "intensity" giving rise to F_1, i.e. F associates observed intensities with points on the object Σ. (In what follows, we will assume that a change in viewpoint does not alter this association, i.e. that the intensity we observe behaves as if it were an intrinsic surface characteristic. This simplification is justified when the changes in viewpoint we will be

considering lead to negligible changes in the intensity associated with a given point on the object being viewed.) This is just the standard imaging situation, slightly generalized.

To be more precise, we consider some surface embedded in $\mathbf{R}^3$ as the 2-manifold Σ embedded via the injection i. Let $\pi : \mathbf{R}^3 \to \mathbf{R}^2$ be the standard projection onto the first 2 factors, i.e., $\pi : (x, y, z) \mapsto (x, y)$, also called *orthographic projection. Perspective projection* can be defined as a map

$$\pi : (x, y, z) \mapsto \left(\frac{x}{kz+1}, \frac{y}{kz+1} \right).$$

Since this map has a singularity at $z = -1/k$, it is not defined on all of $\mathbf{R}^3$. Thus to subsume perspective projection, we have to generalize our picture slightly (but really without changing the essence), as shown in Fig($\frac{s'}{2}$).

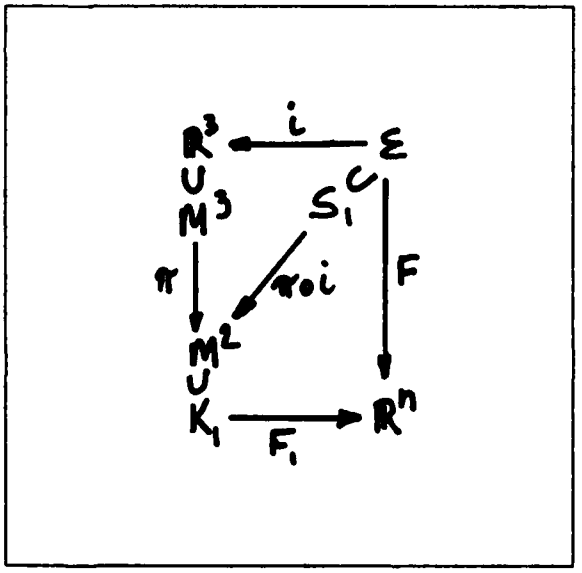


Fig. ($\frac{s'}{2}$)

M^3 is a fixed 3-dimensional subset of $\mathbf{R}^3$, and is the domain of definition for the imaging projection π, which maps it to M^2, the 2-dimensional image space. We require further the physically obvious regularity condition that π be a C^∞ *submersion*, i.e. that its

derivative be everywhere surjective. Usually, $M^2 \subset M^3$ and π is a projection in the sense that $\pi^2 = \pi$. These conditions are all satisfied for ray optics if the rays do not intersect in the image (as might happen if there were caustics, e.g.), i.e. if an image point corresponds to a unique ray. For orthographic projection, $M^3 = \mathbf{R}^3$ and $M^2 = \mathbf{R}^2$. Alternatively, so as not to look in front and behind at the same time, M^3 could be the upper half space of $\mathbf{R}^3$. For the usual perspective projection geometry given above, M^3 can be taken to be the same upper half space. In this case, the singular plane (containing the pinhole) is behind the film plane, $M^2 = \mathbf{R}^2$. Another imaging geometry is interesting at least theoretically, which we call spherical perspective projection. In this geometry, the projection can be looked at in spherical coordinates as the map $\pi : (r, \theta, \varphi) \mapsto (\theta, \varphi)$. With our conventions, $M^3 = \mathbf{R}^3 - 0$, $M^2 = S^2$, the unit sphere, and π is projection onto the sphere along the line to the center, 0. In this case, it is easy to see, e.g. that the space of orientations of the camera is isomorphic to the rotations of the unit sphere.

S_1 and K_1 are corresponding visible regions of Σ and the image space M^2, resp. More precisely, let $S_1 \subset \Sigma$ be such that $i(S_1) \subset M^3$ and $\pi : i(S_1) \to M^2$ is 1-1. Then let $K_1 = \pi \circ i\,(S_1)$. This makes all the pictured maps well-defined, and the diagram Fig. $(\frac{*'}{2})$ commutative.

We assume that the surface Σ admits a function $F : \Sigma \to \mathbf{R}^n$ which describes intrinsic surface features. E.g., in the situation $F : \Sigma \to \mathbf{R}^1$ (i.e. $n = 1$), F can be thought of as representing an intrinsic surface brightness or luminance. Thus we are presently ignoring the effect of viewpoint on image irradiance, or, put another way, we are taking the reflectance function to be constant. To the extent that we deal only with small changes in viewpoint, that will usually be a good approximation. One can enlarge the analysis to include an image irradiance equation, but only with added complexity, so we do not consider this here. If one wishes, F can be thought of as the intrinsic surface property *albedo*, and assume that our analysis deals with quantities that depend only on albedo, to good approximation. For the case $n \geq 2$, we have in mind color images:

normal human cone vision incorporates a function $F_1 : K_1 \to \mathbf{R}^3$ $(n = 3)$. Note that we also subsume cases for a smaller (i.e. $n = 2$) or larger ($4 \leq n < \infty$) number of passbands, or in fact any surface attribute, such as a multi-dimensional texture measure, which can be thought of as taking pointwise values in some real vector space.

We now make precise the imaging geometry which gives us the observed image F_1 from the intrinsic surface function F. Basically, we want to say that we see the frontmost surface of Σ (given i and π, i.e.). This may not take up all of the image plane, and e.g. if Σ is compact then its picture, $\pi \circ i(\Sigma)$ will also be compact. Since π is a submersion, $\pi^{-1}(p)$ (for $_1 \in M^2$) is always locally 1-dimensional. We assume that our imaging projection is sufficiently simple that $\pi^{-1}(p)$ is not a circle; this is true if we assume light travels in straight lines, e.g.

Here is an example of a map $\pi : M^3 \to M^2$ which conforms to all the requirements we have made until now, but for which $\pi^{-1}(p)$ is a circle. Let M^3 be the solid torus $S^1 \times D^2$, where D^2 is the unit disk. Let π simply be projection onto the 2nd factor, i.e. $\pi(\theta, p) \mapsto p$. The situation is illustrated in Fig. (torus)

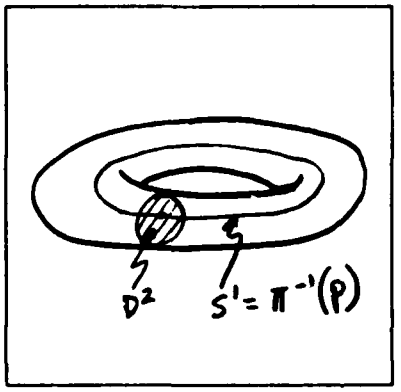


Fig.(torus)

All our regularity assumptions are clearly satisfied. And $\pi^{-1}(p) \approx S^1$.

We assume further that M^3 and M^2 can be embedded in a product structure such that π is projection on one of the factors (we have already assumed that the other factor is a subset of the line). I.e., we assume there is some manifold A and embeddings e_1, e_2 which make the following diagram commutative:

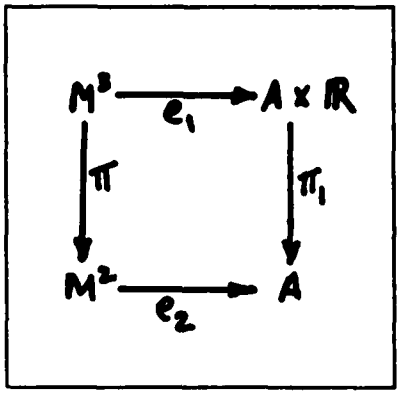


Fig.(prod)

Since we have excluded the circle, $\pi^{-1}(p)$ has a well defined order induced by the usual ordering on the line, so we can define the closest point of some set on any ray as a least element, so long as the intersection with the ray is closed. In fact it is, since the singular set is closed by virtue of being the inverse image of a closed set, while the inverse image of a regular value is closed, since it is a submanifold. Using regularity, the ordering can be extended to a submanifold (with boundary). (The underlying theory is presented later.) Incidentally, the singular set of $\pi \circ i$ is also called the *silhouette* of Σ, since it comprises the points of tangency of the line of sight to the embedding of Σ.

We are now ready to discuss the more involved situation of Fig. (*). For the same reasons that we used Fig. ($*'_2$), we will replace Fig. (*) with Fig. (*'):

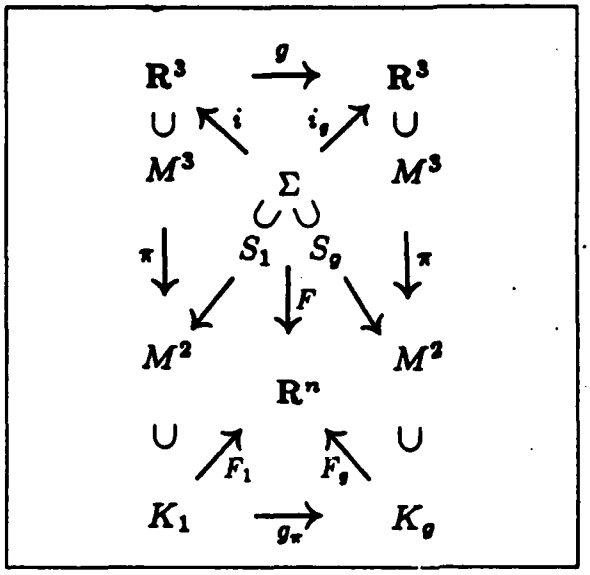


Fig. (*')

The new feature in this picture (beyond Fig. $(\frac{*'}{2})$) is the effect of change in viewpoint. A change in viewpoint means that the imaging projection π changes. Let π_0 be the projection for viewpoint v_0 and π_1 that for v_1, where we loosely define a viewpoint as a location, direction, and orientation (we might tilt our head) of looking. Then π_1 is just π_0 preceded by a change of coordinates. I.e., $\pi_1 = \pi_0 \circ \psi$, $\psi : \mathbf{R}^3 \to \mathbf{R}^3$. In other words we can describe the change in π using an egocentric view where π is constant, but the world moves. The world-o-centric picture is:

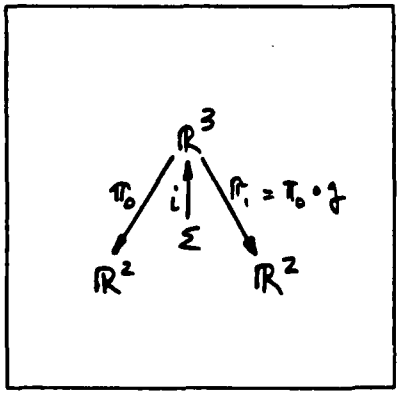


Fig. (world-o-centric)

This is really just a contraction of the egocentric Fig. (*), so we will use the egocentric model, since it is easier to make things explicit that way. The map $g : \mathbf{R}^3 \to \mathbf{R}^3$ is the coordinate change in $\mathbf{R}^3$ (the ambient space in which our objects are embedded). In fact, since we are restricting ourselves only to coordinate changes resulting from a change in viewpoint, we do not want to alter any metric properties, i.e. we want to preserve geometry, so a little thought should persuade one that the possible coordinate changes g are only the orientation preserving isometries of $\mathbf{R}^3$, also known as the rigid motions of $\mathbf{R}^3$, or the Euclidean group $E(3)$. After we have applied the motion g, the new embedding of Σ is given by $i_g = g \circ i$; this just says that the embedded surface got carried along with the motion.

We now have to take some care with the definitions of K_1, K_g, S_1, and S_g. If we were to use just the definitions of 2 copies of Fig. ($\frac{s'}{2}$) pasted together, g_π might not be well defined, since we could not be sure that $S_1 \subset S_g$, or equivalently that $\pi \circ i_g : \Sigma \to M^2$ is 1-1 on S_1, since for many surfaces Σ, different viewpoints g have different domains of visibility of Σ. This is a fancy way of saying that part of what we saw in the picture F_1 might be hidden from view when we look after doing g. Hence the regions K_1, K_g must be chosen in such a way that g_π is well-defined. For example, having chosen S_1, S_g as above, we can define $S_1' = S_g' = S_1 \cap S_g$ and $K_1' = \pi \circ i(S_1')$ and $K_g' = \pi \circ i_g(S_g')$. With these restrictions, g_π is a diffeomorphism $K_1 \to K_g$ with the property that $F_1(p) = F_g(q) = F_g(g_\pi(p))$, which is the same as saying that g_π is a deformation of the picture F_1 into the picture F_g. Note that this observation is also equivalent to asserting that the diagram is commutative (for the F_1, g_π, F_g loop).

Occlusion, the obscuration of one part of surface by another, occurs at the singularities of the mappings $\pi \circ i$ and $\pi \circ i_g$. Self-occlusion can occur for many objects, and if we allow Σ to have more than 1 connected component, we are able to subsume all cases of occlusion. The complicating feature then becomes that the domains of smoothness of F_1, F_g are bounded by the singular sets, and an important problem then is to understand the

singularities. We have not considered this problem here, but the topology involved has been well-studied in singularity theory and catastrophe theory (see [Arnold 1984, Arnold, V.I. 1983] for expositions of the theory by one of the grandmasters, and [Koenderink and van Doorn 1982, Koenderink and van Doorn 1976] for some discussion in the context of vision).

A Catalog of Applications

Now that we have established an abstract description of the mathematical setting for vision, we can indicate how the usual computer vision problems fit into the structure.

We consider these problems:

- Area matching stereo
- General matching
- Motion stereo and optical flow
- Feature based stereo
- Singularity tracking

In subsequent sections, we will prove theorems about general matching and optical flow. The structure we have presented comprises differentiable mappings among various spaces, and, emulating the *Erlanger Programm*, a group - rigid motions in 3-space. The mathematics of these structures is differential topology and differential geometry, so we turn to the tools of these trades for our analysis.

For general matching, the central result is that ***unique image matching requires at least 2 color dimensions***, unless one has knowledge of imaging geometry. ([Resnikoff 1974] studied some relations between color and geometry, but in a quite different context.) The results for optical flow show how to exploit this knowledge, using the geometry of rigid motions of 3-space in the form of Lie group theory. We will also discuss the topological structure of images, and show how the invariants can be used for stereo matching as well as image understanding via "scale space" [Witkin 1983].

Stereo

For the stereo problem, we assume that we are given 2 pictures F_1, F_g arising from simultaneous views of a surface in $\mathbf{R}^3$. There may be some constraint, or even complete knowledge, of the viewing situations which gave rise to the images, i.e. we may have information about the camera model. We want to find the topography of the surface.

It should be evident that the situation is exactly that of Fig. (*), with the restriction that K_1 and K_g are projections from the single surface Σ (we could moot the restriction by allowing Σ to have more than 1 component; however that creates complications, and we consider the simpler case). Then

Given

F_1, F_g	pictures

We want to find

i	surface embedding
g_π	picture correspondence
F	intrinsic surface characteristic

Examples of possible viewpoint constraints are:

1) $g \in E(3)$ given	camera model completely specified
2) $g = g_t$ for some $t \in \mathbf{R}$, where g_t is defined by $\gamma : \mathbf{R} \to E(3)$, $g_t(p) = \gamma(t)(p), \quad p \in \mathbf{R}^3$	viewpoints on a 1-parameter subset of $E(3)$
3) in addition, $g_{t+s} = g_t \circ g_s$	viewpoints on a 1-parameter subgroup of $E(3)$
4) g_t leaves each of a space-filling set of parallel planes invariant for each t	viewpoints are combinations of translations and rotations such that epipolar lines are well-defined

a point at which all the partials of f vanish. The x_i here are *coordinate functions* on M^m, defined for e patch φ (we omit the precise definition, which can be found in any differentiable manifold book).

w the only boundaryless 1-manifolds are the diffeomorphs of $\mathbf{R}^1$ and S^1 (the circle) lnor 1965], so in a region where f has only regular values, our picture is essentially rect. I.e., the level set corresponding to a regular value must be a 1-manifold. Now need to know that almost all values are regular; then since each value determines a el set, almost all level sets will be 1-manifolds as we are claiming. But so far, we don't n know that there has to be *any* region (i.e. neighborhood) free of critical values. In , if F_1 is a constant map, then clearly all of U_1 consists of critical points.

eorem (Sard) Let $f : M^m \to M^n$ be a C^k mapping between the m, n-dimensional nifolds M^m, M^n, where $k > \max(m - n, 0)$ (for a monochrome picture this means > 0, i.e. f is differentiable). Then the Lebesgue measure of the set of critical *values* M^n) is 0.

lnition In a measure space, ***almost all*** means all but a set of measure 0.

nark In a probability space, ***almost all*** is equivalent to ***with probability 1.***

d's theorem says, in other words, that almost all values are regular, which is the same claims 1) and 2) above. Note that it is the critical *values* that are of measure 0, not critical *points*. Thus, for us, this means that the set of intensity values (but not essarily picture points) taken at critical points (where a level set is not a 1-manifold) parse. It could still be dense however, e.g. if there were critical values at all the onals.

Typically, pictures have isolated critical points (i.e. they do not form blobs, lines, or umulations)

tunately, we can say more. There are certain nasty types of critical points called *enerate* and nice ones called *nondegenerate* (we'll define them in a moment). One he nice things about nondegenerate critical points is that they are *isolated*, i.e. a degenerate critical point has some neighborhood which contains no other critical

The *Jacobian* of f is really defined with respect to some pair of coordinate systems on M^m and M^n. Let a *patch* of the manifold M^m be a smooth 1-1 map $\varphi : U^m \to M^m$, where $U^m \subset \mathbf{R}^m$ is a neighborhood in $\mathbf{R}^m$, such that $\varphi^{-1}|_{\varphi(U^m)}$ is also smooth. (This is just like our definition of a *patch* of a surface earlier.) Suppose we're interested in the Jacobian at a point $x \in M^m$. Then let ψ be a similarly defined patch in M^n such that $f(x) \in \psi(U^n)$, i.e. so that f puts x into the right region for ψ. Then $\psi^{-1} \circ f \circ \varphi$ is a map $\mathbf{R}^m \to \mathbf{R}^n$, and we can speak of the classical Jacobian of this map, defined as follows. The *Jacobian matrix* of $F : \mathbf{R}^m \to \mathbf{R}^n$ at p is the matrix of derivatives $\partial F_i / \partial x_j$, i.e.

$$J_p(F) = \left[\frac{\partial F_i}{\partial x_j}(p) \right]$$

Although the Jacobian itself depends on the coordinate systems chosen for M^m and M^n, its rank does not (see e.g. [Golubitsky and Guillemin 1973]). Thus we can speak of the rank of the Jacobian of f above, even though by definition we just presented, the Jacobian itself depends on the particular coordinate charts. One can also give a coordinate-free definition of Jacobian, where the Jacobian of f is the *derivative* of f, a map between tangent spaces, and then the Jacobian of f is a unique, well-defined object. The interested reader can find the details in any book explaining differentiable manifolds, e.g. [Abraham and Marsden 1978, Golubitsky and Guillemin 1973, Guillemin and Pollack 1974, Hirsch 1976]. The Jacobian is nothing more than the linear approximation to the map; or it can be thought of as the linear term in the Taylor series, which is the same thing. Thus it gives information on what the map does to the degrees of freedom in the domain space.

Here are some related definitions:

Definition. Let $f : M^m \to M^n$, be C^1.

1) $p \in M^m$ is a *regular point* of f if the Jacobian of f at p is of maximal rank.

2) $p \in M^m$ is a *critical point* of f if it is not a regular point, i.e. if the rank of the Jacobian of f at p is less than maximal.

3) If $p \in M^m$ is a critical point of f, then $f(p) \in M^n$ is a *critical value* of f. Note this means that $q \in M^n$ is a critical value of f if $f^{-1}(q)$ contains a critical point, even though it may be that $q = f(p')$ for some regular point p. Also notice that mountain peak *heights* are critical values.

4) $q \in M^n$ is a *regular value* of f if it is not a critical value. So q is a regular value if $f^{-1}(q)$ contains only regular points, *or* if q is not even in the range of f. That's because it's handy to have only 2 types of points in M^n: critical and regular.

5) f is an *immersion* at p if p is a regular point and $\dim M^m \leq \dim M^n$.

6) f is a *submersion* at p if p is a regular point and $\dim M^m \geq \dim M^n$.

7) If f is an immersion (submersion) at every $p \in M^m$, then it is simply called an *immersion* (*submersion*).

8) f is an *embedding* if it is an immersion and a homeomorphism onto its image. [A *homeomorphism* is a mapping which is continuous and has a continuous inverse.]

There are numerous versions of the implicit function theorem, which go by various names, the most common of which is the inverse function theorem. The above version is one of the most general. The theorem is frequently stated only for the case $m \geq n$, and the condition may be stated in terms of regularity, rank or singularity (as a matrix or linear map) of the Jacobian or derivative, *transversality* of f, etc. What all these essentially mean is that at the point in question, f only does as much collapsing as is required to squeeze things into the dimension of the range, and no more. Notice that mountain peak *heights* are critical values.

In our case, we are currently dealing with the situation of 1 color dimension, so we are interested in $F_1, F_2 : M^2 \to \mathbf{R}^1$. Thus the theorem tells us that for a regular value y of F_1 (resp. F_2), ${F_1}^{-1}(y)$ (resp. ${F_2}^{-1}(y)$) is a 1-dimensional submanifold of $K_1 = M^2$. Note that for a function $f : M^m \to \mathbf{R}^1$ the Jacobian is an $n \times 1$ matrix, so a critical point p of f is one for which

$$\frac{\partial f}{\partial x_1}(p) = \cdots = \frac{\partial f}{\partial x_n}(p) = 0$$

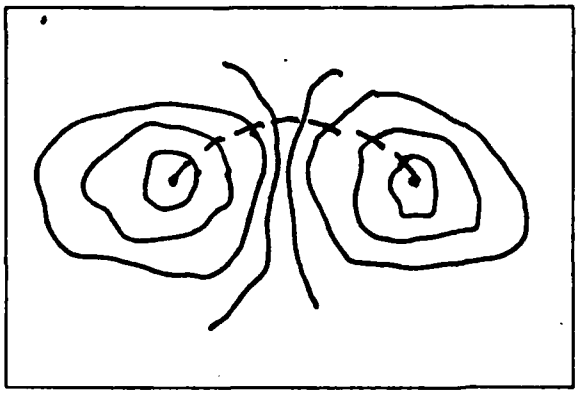


Fig. (frag)

Here is the gist of what we will say in more precise terms.

1) Almost every level set of a picture is a circle or a line

2) These 1-manifolds account for almost all of the brightness values; the rest are extrema or saddles (critical points).

3) Typically, pictures have isolated critical points (i.e. the critical points do not form blobs, lines, or accumulations).

1) and 2) Almost all level sets and brightness values are regular

First we need to know that the contour lines have the simple structure above. To this end we need the following version of the

Implicit function theorem (see e.g., [Bröcker and Lander 1975, Nitecki 1971, Golubitsky and Guillemin 1973]). Let $f : M^m \to M^n$, be C^r. (Where M^k denotes some k-dimensional manifold.) Then $f^{-1}(y) \subset M^m$ is a C^r submanifold of dimension $\max(m-n, 0)$ (or empty) if the Jacobian of f is of maximal rank (i.e. rank $\min(m, n)$) at each $x \in f^{-1}(y)$.

Note that $f^{-1}(y)$ cannot be self-intersecting, since it is a submanifold.

differentiable 1-dimensional objects. That in turn is a consequence of the fact that the picture is a map from a 2-dimensional object to a 1-dimensional object.

Some differential topology for vision

In the proof of the first part of the theorem the facts we used from differential topology, about gradient vector fields, flows, and contour lines, were elementary enough that we did not have to go into much detail about the theory behind them. For higher dimensions, in particular for more color dimensions, the situation is more difficult; and several advanced ideas are prerequisite to proceeding with the rest of the proof. Also, the proof we just gave for the monochrome case is somewhat technical, so we would like to illuminate the intuitive ideas with some deeper results. The rest of this section, therefore, is devoted to a review of some of the necessary ideas of differential topology, integrated with establishing (for the first time in the vision literature) the aspects of vision to which they correspond. We will use this theory in later sections as well; moreover, it is basic to the geometric aspects of vision.

First, let's see how the proof given above fits into the intuitive scheme presented earlier for using Fig. (frag). Then we will worry whether Fig. (frag) is a reasonable picture for the contour lines of a picture function. The φ_ρ which we defined above, considered along the dotted line, is essentially the rotation function we discussed earlier. We could use a bump function to make it just what we want in some neighborhood of the dotted line, and shear could be eliminated by using a ρ with negative as well as positive values. It would take a little work, but it could be made to do the right thing.

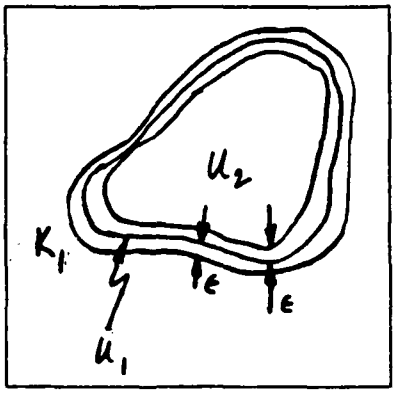


Fig. (boundary)

By a standard construction (e.g. [Abraham and Marsden 1978]), there is a C^∞ function $K_1 \to \mathbf{R}$ which takes the value 0 outside of U_1 and the value 1 on U_2. Using this "bump" function $\beta : K_1 \to \mathbf{R}$, we get a vector field $\beta \cdot Z$ on K_1 which vanishes outside of U_1, hence its flow never leaves K_1, i.e. the flow φ_t of the vector field $\beta \cdot Z$ has the property that $\varphi_t(p)$ is defined and lies in K_1 for all $p \in K_1$ and $-\infty < t < \infty$. Hence for any such t, $\varphi_t : K_1 \to K_1$ constitutes a diffeomorphism with contour lines invariant on K_1. In fact, it is easy to see that this family of diffeomorphisms can be enlarged even more. Notice that multiplying the vector field Z by a scalar C^r function $\rho : K_1 \to \mathbf{R}$ does not alter orbits. We can therefore enlarge the class of diffeomorphisms φ_t by taking all diffeomorphisms $\varphi_{t,\rho}$ given by the flows of $\rho \cdot \beta \cdot Z$ on K_1. Observe that for any constant α, $\varphi_{\alpha t,\rho} = \varphi_{t,\alpha\rho t}$, so if ρ is a constant function, $\varphi_{t,\rho} = \varphi_{\rho t,1} = \varphi_{1,\rho t}$. Thus $\{\varphi_{t,\rho}\} = \{\varphi_{1,\rho}\}$, so by abuse of notation we will write φ_ρ for $\varphi_{1,\rho}$. **QED** $(n = 1)$.

We have thus far proved our result for the monochrome case. In summary, we have shown that in matching 2 regions free of occlusions, i.e. when we have a matching diffeomorphism between the regions, the match is far from unique. In fact there are essentially as many matches as C^r functions from such a region to the reals. (One has to factor out functions ρ which lead to equivalent time-one maps, e.g. those which rotate contour lines by multiples of 2π.) This stems directly from the fact that the iso-brightness loci constitute connected

inner product on an orientable manifold). One might, e.g. define the new vector field Z on K_1 by $Z(p) = (-b, a)$ if $\nabla F_1(p) = (a, b)$. Note that $Z \cdot \nabla F_1 = 0$ at all p. Since smoothness is defined with respect to coordinates, Z has the same degree of smoothness as ∇F_1. Furthermore, wherever $Z \neq 0$, it is tangent to the contour lines of F_1, so that the orbits of Z are exactly those contour lines, and the critical points are exactly the critical points of ∇F_1. We now want to consider the flow $\varphi_t : K_1 \to K_1$ of the vector field Z.

The *flow* $\varphi_t : M \to M$ of a vector field on the space M is the solution to the initial value problem defined by considering the vector field as a system of differential equations on M. I.e., φ_t is the unique map such that $d/dt\, \varphi_t(p) = v(p)$, where $p \in M$ and $v(p)$ is the vector at p. The flow moves the space along the solution lines, which are always tangent to the vector field. Smoothness of the vector field guarantees smoothness (and uniqueness) of the flow. The *time-one* map associated with a flow φ_t is the diffeomorphism $\varphi_1 : M \to M$; i.e. a snapshot of the flow at one particular instant of time. The *orbit* of a point (or set) p under the flow, is the set of all values of $\varphi_t(p)$, for all t, i.e. $-\infty < t < \infty$. Notice that the flow is a function of time as well as a map on the manifold; this is a slight abuse of the notation we are using for functions.

But first we have to deal with a slight problem, viz. that near the boundary of K_1, the time-one map may not be defined if a contour line has a boundary. To overcome this, we use the following device to make Z vanish near the boundary of K_1. We find open sets U_1, U_2 such that $\overline{U}_2 \subset U_1 \subset \overline{U}_1 \subset K_1$. U_2 can be almost as big as K_1, since we can choose U_1 and U_2 as follows. Let $\overline{U}_1$ be $K_1 - V_\epsilon$, where V_ϵ is an ϵ -neighborhood of the boundary of K_1, where by ϵ -neighborhood we mean the union of all open balls of radius ϵ centered at points of the boundary of K_1. Then, of course, $U_1 = \text{interior}\, \overline{U}_1$. Similarly, U_1 can be slightly contracted to yield U_2. If we assume that the boundary of K_1 is piecewise smooth (which follows, e.g. if the picture results from a finite number of smooth objects) then the measure of U_2 can be made arbitrarily close to that of K_1.

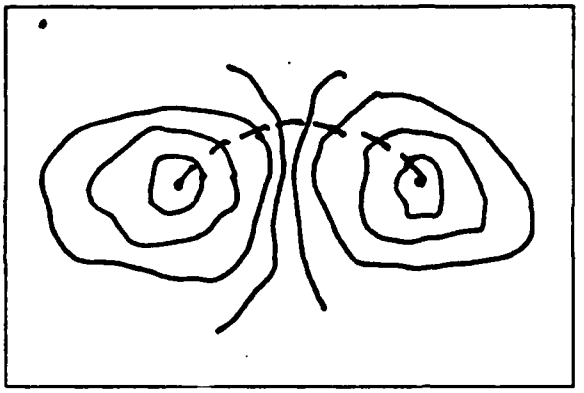


Fig. (frag)

Observe first that if $\psi : K_1 \to K_1$ is a diffeomorphism taking contours of F_1 to contours of F_1, then g_π is a matching function $\Rightarrow g_\pi \circ \psi$ is a matching function. Define ψ as follows. As you go along the dotted line

$$\begin{aligned} \gamma : I &\to K_1, \qquad I = [0,1] \subset \mathbf{R} \\ t &\mapsto \gamma(t) \end{aligned}$$

in Fig (frag), slide each contour along itself by an amount $\theta(t)$. As long as $\theta : I \to \mathbf{R}$ is a diffeomorphism onto its image, the map ψ will be a diffeomorphism in a neighborhood of the dotted line. To the extent that this picture is valid, there will be as many matchings $g_\pi \circ \psi$ as there are such maps θ.

Actually, we are going to use a slightly more general method to construct a family of diffeomorphisms ψ_α roughly in 1-1 correspondence with the set of all C^r functions $K_1 \to \mathbf{R}$. For this we will use a canonical vector field defined along the contour lines of F_1, which will tell us how much to slide each contour line.

First we observe that the map $F_1 : K_1 \to \mathbf{R}$ has a canonical vector field associated with it, the gradient vector field ∇F_1 which assigns to each point $p \in K_1$ a vector $\nabla F_1(p)$. ∇F_1 is always orthogonal to the contour lines of F_1 (with the usual inner product on K_1 inherited from $\mathbf{R}^2$), and it is 0 precisely at the critical points of F_1. In (2-dimensional) coordinates, $\nabla f = (\partial f/\partial x, \partial f/\partial y)$. Clearly, if $f \in C^r$, then $\nabla f \in C^{r-1}$.

Define a new vector field on K_1 by rotating each of the local vectors of ∇F_1 by $+90°$, i.e. $+90°$ counterclockwise, (which is uniquely defined because we have a globally defined

If $n \geq 3$ (i.e. the picture has at least 3 color dimensions), then generically there will be a unique g_π which makes the diagram commute.

Once the problem is appropriately formulated, the proof yields to repeated attack by some standard machinery of differential topology. (An excellent introduction to the subject is [Guillemin and Pollack 1974], and [Hirsch 1976] is a good reference.)

Proof (case $n = 1$). For the time being, we only consider monochrome pictures ($n = 1$). We will return later to the situation for pictures with more color dimensions.

The idea for this part of the proof is fairly simple; the difficulty lies in establishing when it is valid.

The map $F_1 : K_1 \to \mathbf{R}$ can be thought of as a topographic landscape on $K_1 \subset \mathbf{R}^2$, where intensity is represented by altitude. Consider $K(x) = {F_1}^{-1}(x)$ for some intensity $x \in \mathbf{R}$. $K(x)$ is an iso-intensity contour for the intensity x and corresponds to an elevation contour on a geographic topographic map.

The idea is this. Observe that if $F_2 \circ g_\pi(p) = F_1(p)$ (i.e. if Fig (α) is a commutative diagram) then $g_\pi({F_1}^{-1}(x)) = {F_2}^{-1}(x)$, i.e. g_π takes contour lines to contour lines. (Proof: Suppose $p \in {F_1}^{-1}(x)$ and $q = g_\pi(p)$. Since $F_2(q) = F_1(p)$ and $F_1(p) = x$, $q \in {F_2}^{-1}(x)$. Thus $g_\pi({F_1}^{-1}(x)) \subset {F_2}^{-1}(x)$. Similarly, since g_π is a diffeomorphism, $(g_\pi)^{-1}({F_2}^{-1}(x)) \subset {F_1}^{-1}(x)$ whence ${F_2}^{-1}(x) \subset g_\pi({F_1}^{-1}(x))$.) Conversely, *any* diffeomorphism $h : K_1 \to K_2$ which takes contour lines of F_1 to contour lines of F_2 satisfies the conditions for g_π. (Proof: Essentially immediate: We want to show that $h({F_1}^{-1}(x)) = {F_2}^{-1}(x) \Rightarrow F_2(h(p)) = F_1(p)$. Choose $p \in K_1$, and let $x = F_1(p)$ so that $p \in {F_1}^{-1}(x)$. By hypothesis, $h(p) \in {F_2}^{-1}(x)$, so $F_2(h(p)) = x = F_1(p)$. QED)

Thus far we have shown that any g_π taking contour lines to contour lines will solve our local matching problem. But how many such g_π's can there be? Assume for the moment that a typical contour map contains a diffeomorphic image of the fragment represented by the solid lines below

confine our attention only to a diffeomorphism $g_\pi : K_1 \to K_2$ where K_1, K_2 are both connected.

The 2-color theorem

Theorem (2-color theorem). Stereo requires at least 2 colors or 3 dimensions. I.e., for a monochrome picture, general matching has infinitely many solutions, but for 2 or more color dimensions, it is generally unique. Hence the monochrome case requires knowledge of the imaging situation to constrain the problem.

More precisely, consider the commutative diagram

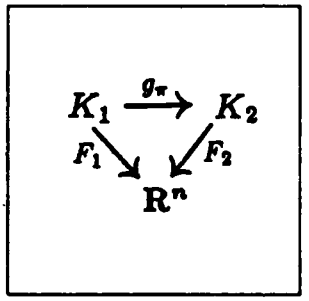


Fig. (α)

where g_π is a C^1 diffeomorphism , F_1, F_2 are C^1, and K_1, K_2 are compact.

If $n = 1$ (i.e. the picture is monochrome), then $\exists$ an infinite-dimensional family of C^1 diffeomorphisms $\{h_\varphi\}$ such that replacing g_π by h_φ also results in a commutative diagram (i.e. is a solution). The family h_φ is parametrized by (at least) the continuous functions $K_1 \to \mathbf{R}$, and contains an isomorph of a neighborhood of the identity.

If $n = 2$ (i.e. the picture has 2 color dimensions), then generically there will be a finite number of g_π which make the diagram commute (note we have assumed that such a g_π exists). If we take K_1, K_2 to be rectangles or discs (as in a usual picture) then generically there is a unique g_π.

al. 1983]). The epipolar geometry consists of known foliations of K_1 and K_2 by curves, with the image of each such curve under g_π also known. This information is derived from sources other than general matching; e.g. from singularity matching, or interactive (i.e., human) guidance, used along with assumptions about imaging geometry, such as optical characteristics. Part of our purpose is to understand how this information provides a constraint, and ultimately to see how much of it can be found in an integrated process.

We are not proposing here to use exact equality of point brightness values as a matching criterion for stereo vision programs, nor to ignore imaging geometry. Rather, we are investigating the consequences of the idea that there is *some* function describing surface character, maybe not the data itself, which manifests itself in 2 different distorted pictures. We want to know what it takes to find that distortion *in principle*. We see this as a first step to understanding what it takes *in practice*, where there are further complicating factors. We consider some of these in later sections.

The following question then arises:

Problem (Uniqueness of General Matching). If we are looking for an arbitrary (piecewise) C^1 diffeomorphism g_π to make Fig.(GM) commute, under what conditions are we guaranteed a unique solution to the matching problem?

E.g., if F_1, F_2 are both constant functions, i.e., we have uniformly gray pictures, the problem is completely degenerate, and any diffeomorphism g_π is a solution.

Since there may be occlusion, K_1 and K_2 may not be connected regions. We don't consider the (important) problem of determining the connected components of K_1 and K_2, i.e. determining the occlusion-free regions. Suppose, instead, that some g_π exists fulfilling the above criteria. We will be concerned only with regions where g_π is smooth, i.e., a C^1 diffeomorphism. These are regions containing no occlusions or points where the surface is tangent to the line of sight. Furthermore, we do not consider the problem of determining which are the corresponding connected components of the 2 pictures. We

This means that every point in one region is matched to a corresponding one in the other, in such a way that the 2 picture functions give identical brightness or color values on corresponding points, while keeping the distortion continuous and differentiable, i.e. maintaining the region topology. This automatically guarantees matching of context. Only after the matching function is found is the surface embedding computed by associating relative depth with relative disparity at each point of (say) K_1. In this approach the matching proceeds without any knowledge of the 3-dimensional structure represented by Fig. (*). Stereo matching programs rarely actually try to solve the problem in this pure form, for a number of good reasons. In the first place, geometric information *is* usually available, and some of it is often used to constrain the matching. In fact, as we will show, this is necessary to achieve any success for unique point correspondence. Secondly, programs do not usually use simple equality of brightness values as a matching criterion (though see [Baker 1981] for a use of essentially that criterion as an interpolation method for regions between known corresponding points). There are several reasons for this. Various sources of noise, including digitization as well as electronics, make it impractical to look for exact values of brightness. There can be variations between 2 images, such as camera settings or film properties, as well as photometric changes. In approaches which use area matching (e.g. [Gennery 1980]) one frequently uses some measure of similarity of context as a matching criterion; one family of these is derived from cross-correlation. Part of the art of these measures is to compensate for the imperfections we have just mentioned. Nevertheless, there is generally the assumption that there is some underlying function which transforms according to Fig.(GM); although this function may not be identical with the data, it gives rise to it.

Frequently one assumes that the 2 images F_1, F_2 are *rectified*, i.e. that g_π takes scan lines to scan lines in a known way: $g_\pi(x,y) = (\tilde{g}(x,y), y)$ for some $\tilde{g} : K_1 \to \mathbf{R}$. This very strong constraint on imaging geometry is rarely valid in practice. Instead, one can rely on knowledge of *epipolar geometry* for an additional constraint (see, e.g., [Baker et

An Application: Stereo by General Matching

As an application of the abstract viewpoint we are proposing, we show that for monochrome images, the general matching problem is insoluble. We exhibit the degeneracy, and show that additional color dimensions allow unique solution.

The problem

A common goal of stereo matching is to solve the *correspondence problem* for some region, i.e. to pair corresponding points between 2 pictures within some region. A pair of points in 2 pictures *correspond* if they arise from a common single point in the scene. The correspondence must be inferred from the picture functions. There have been many approaches taken to do this, and geometric information as well as a point's picture context have been used in many ways to make the inference. One of our ultimate goals is to build a theory which gives a coherent view of the problem and the methods which have been used to attack it.

A basic need is to understand what the roles of geometry and context are in this problem: how much can you tell just from picture distortion, and how much do you have to know about the way the image was formed? To shed some light on this, we look at the stereo problem as the general matching problem. I.e., given 2 picture functions $F_1, F_2 : M^2 \to \mathbf{R}^n$, one finds regions $K_1, K_2 \subset M^2$ and a 1-1 *matching* function $g_\pi : K_1 \to K_2$ such that the diagram Fig. (GM) commutes.

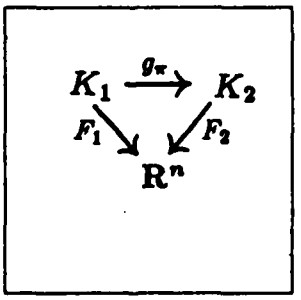


Fig. (GM)

neighborhoods K_i. So, e.g. the assumption of rectified images could be stated as the requirement that g_π take horizontal straight lines to horizontal straight lines.

Motion stereo

Instead of a single $g \in E(3)$, we have a 1-parameter family $\{g_t\}$, given by

$$\gamma : I \to E(3)$$
$$t \mapsto g_t$$

such that $g_0 = 1$ (the identity in $E(3)$), where $I = [0, 1] \subset \mathbf{R}$. Given is a corresponding family of pictures F_t.

It's common to consider a sequence of pictures related by a sequence of transformations $\{g_i \quad i = 0, 1, \ldots\}$, with the corresponding family of pictures F_i. This can be thought of as a special case of the above, where the transformations are parametrized by a discrete set:

$$\gamma : \mathbf{Z}^+ \to E(3)$$
$$i \mapsto g_i$$

Although this reflects the discrete character of what happens in practice, the former (continuous) representation makes it easier to exploit the temporal smoothness properties of the image.

Fig. (*) illustrates the situation for only 2 g_t 's.

5) g_t is a translation for each t the simplest case for epipolar lines

These constraints consist of focusing attention on subsets of $E(3)$ having particular properties.

Area matching stereo

To differentiate from feature-based stereo, we define area matching stereo by requiring that the stereo problem be solved for a full-dimensional part of the surface, i.e. a neighborhood. This bears an implicit assumption that area-supported functions F_1, F_2 are used directly, and that some intrinsic area-supported function F can be found. An example is matching of areas based on the cross-correlation function between the 2 picture functions on those areas. Feature-based stereo, by contrast, depends on lower-dimensional objects, such as edges or *critical points* (precisely defined later).

General matching

This we define to be area matching stereo without any knowledge of imaging, described by the diagram:

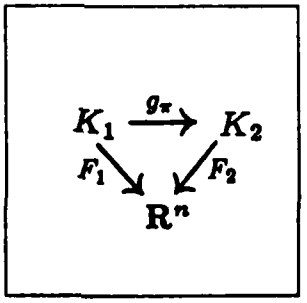


Fig. (GM)

Here we are only given F_1, F_2 and the problem is to find K_1, K_2, g_π such that the diagram commutes. There may be constraints on g_π equivalent to those for stereo, except the constraints can only, of course, be stated in terms of the diffeomorphisms between

points. If we're talking about a region contained in a compact set, that implies a finite number of nondegenerate critical points and a minimum spacing between them. Similarly, there are nice functions which have all their critical points nondegenerate; these are known as *Morse* functions (after Marston Morse). And finally, almost all C^r functions are Morse functions (we'll have to specify what we mean by "almost all"), so that we have a justification for acting as if all our critical points are nondegenerate. Now, here's what all this means in terms of our illustration of level sets of the intensity map (Fig. (frag)). Choose a picture at random. (Say the picture is bounded by a rectangle R with interior V .) If it has no critical points, then all the level sets are diffeomorphic to (disjoint unions of) line segments (and not circles, which are the only other possibility by Milnor's result, cited above). (Proof: By contradiction. Suppose $f^{-1}(a) \subset V$ is diffeomorphic to a circle, so that it bounds a disk in V. The closed disk is compact, so f must take a maximum and minimum on it. If one of these is not on the boundary, $f^{-1}(a)$, it is a critical point of V. If both are on the boundary, then the entire disk consists of critical points, since the maximum and minimum are both a. QED.) Suppose the picture does have critical points. Then "generically" the critical points are isolated.

First let's see what happens near such a critical point. By Morse's Lemma (stated completely below) we know that there is a coordinate system (u, v) in a neighborhood of the critical point p such that $f = f(p) \pm u^2 \pm v^2$. The possible signs correspond to a maximum $(--)$, a minimum $(++)$, and a saddle $(+-)$. So for an extremum, it's easy to see that the level sets are just a point surrounded by circles.

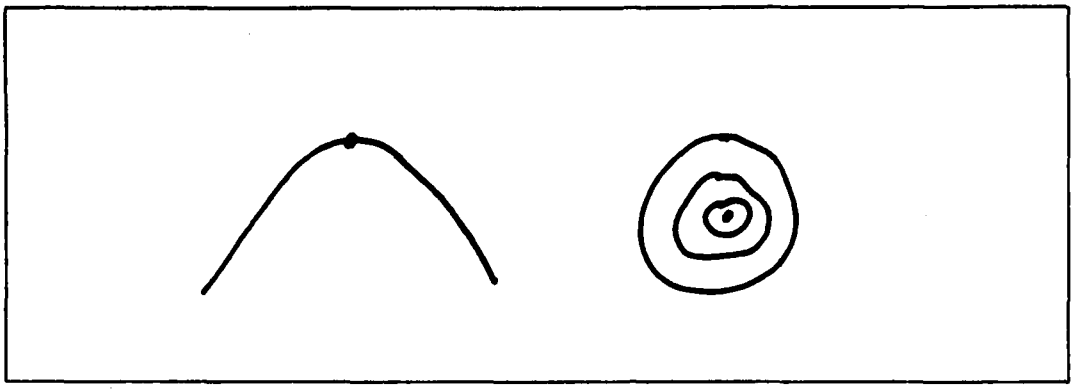


Fig. (extremum)

For a saddle, the level sets are the sets $u^2 - v^2 = const$, which look like

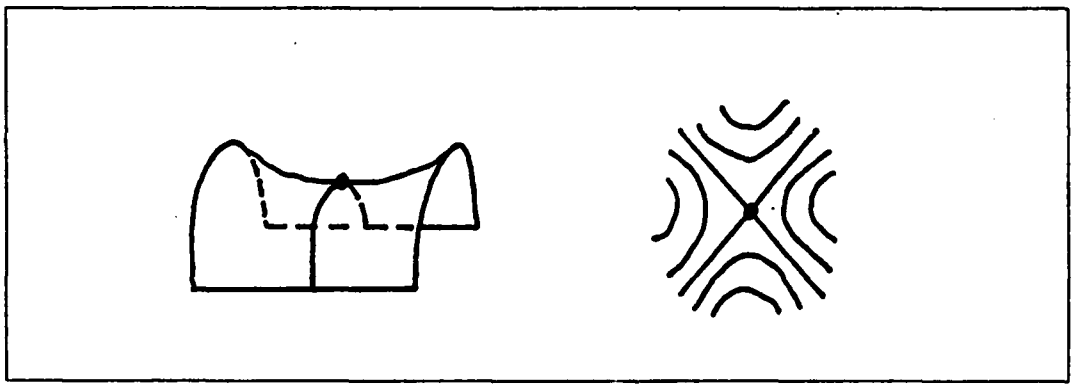


Fig. (saddle)

Note that the critical point is isolated (from other critical points), though it is not isolated as part of a level set.

The Morse inequalities tell us that the Euler characteristic is related to the number and type of critical points. In our case, if we assume that the whole region of interest lies within a single circular level set, this means that the number of extrema must be 1 more than the number of saddles. In Fig. (frag), for the rotation directions of the level sets to be consistent with the way we proved the first part of the theorem, we must assume that

one of the critical points is a maximum and the other a minimum. But from the Morse inequalities, there must be a saddle somewhere, too. In fact, the larger picture looks like.

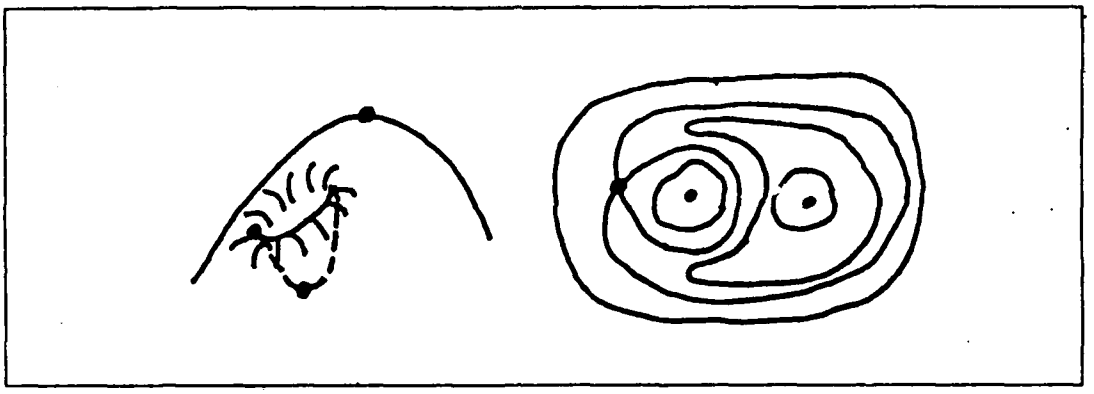


Fig. (dimple)

When there are two maxima (or minima, in Australia), the picture is

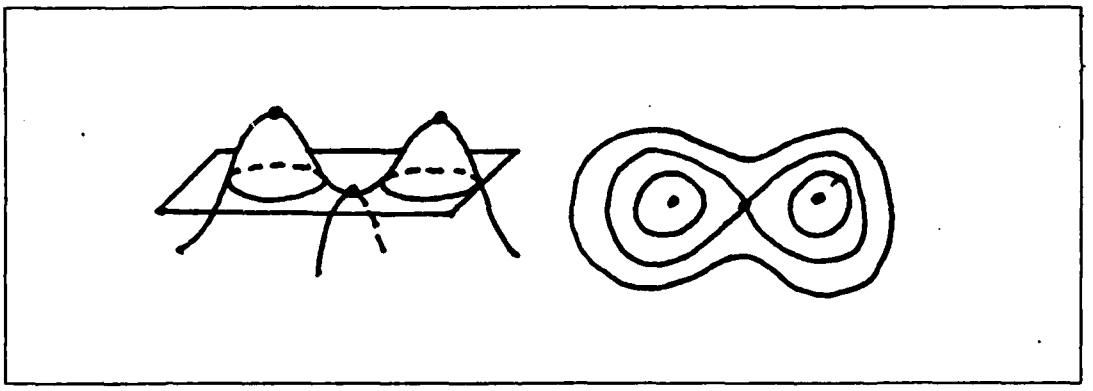


Fig. (pass)

Now we make this precise. [We present the material here in the reverse of the usual order, i.e. we present the main theorem first, and then the definitions required to understand it, since that is the order in which one actually tries to understand the idea.]

Theorem. (see e.g. [Hirsch 1976]) For any manifold M, Morse functions form a dense open set in $C^r_g(M, \mathbf{R}), 2 \leq r \leq \infty$.

Note that in our case, M is the 2-dimensional support of the picture.

Definition. $C^r_S(M,N)$ is the space of all r times continuously differentiable functions $M \to N$, with the so-called *strong* topology. We omit the definition of this topology, but only mention that it is based on the closeness of all derivatives from the 0th (i.e. the value of the function itself) through the rth.

Definition. A *Morse function* is a function $f : M \to \mathbf{R}$ which has only nondegenerate critical points.

Proposition. Nondegenerate critical points are isolated. I.e., a nondegenerate critical point has a neighborhood in which there are no other critical points.

Definition. A *nondegenerate* critical point is one where the Hessian matrix is nonsingular. This basically means that the graph of the function is not flat at the critical point.

Definition. The *Hessian matrix* of a function $g : \mathbf{R}^n \to \mathbf{R}$ at a point p is the matrix

$$\left[\frac{\partial^2 g}{\partial x_i \partial x_j}(p)\right]$$

Theorem (Morse's Lemma). Let $p \in M^n$ be a nondegenerate critical point of index k of a C^{r+2} map $f : M^n \to \mathbf{R}$, with $1 \leq r \leq \omega$. Then there is a C^r chart (φ, U) at p such that

$$f \circ \varphi^{-1}(u_1, \ldots, u_n) = f(p) - \sum_{i=1}^{k} u_i^2 + \sum_{i=k+1}^{n} u_i^2$$

Definition. $p \in M^n$ is a nondegenerate critical point of *index* k of a map $f : M^n \to \mathbf{R}$ if the Hessian of f at p has k negative eigenvalues (counting multiplicities).

Theorem (Corollary of Morse inequalities and Theorem of Hopf). Let $f : M^n \to \mathbf{R}$ be a Morse function on a compact manifold without boundary, with ν_k critical points of index

k, $0 \leq k \leq n$. Then

$$\sum_{k=0}^{n} (-1)^k \nu_k = \chi(M^n),$$

where $\chi(M^n)$ is the Euler characteristic of M^n.

Open dense, usually, generically, almost all, typically

The key result of the theorem above is that the Morse functions are open dense. This allows us to restrict our attention only to pictures whose critical points are isolated and thus to avoid considering pathological behavior.

Here's why. Instead of discussing only pictures and Morse functions, we will talk about dense open subsets of $C^r(M, N)$ generally, since the scope then includes things like general position as well as other properties, with no extra difficulty. Suppose some open dense set consists of functions which all have some nice property. (We will call such a property *generic*. Often generic is defined with respect to a countable intersection of open dense sets, but for us open dense is enough.) Then, as a consequence of density, any function in $C^r(M, N)$ is arbitrarily close to a nice one, hence can be arbitrarily well approximated (with respect to all r derivatives) by a nice one. Of course, we need more than density to be justified in saying "most." Dense sets can have measure 0. For example, both the rationals and irrationals are dense in $\mathbf{R}$, yet we don't want to say that most numbers are rational. Requiring that the set be *open* dense solves this problem (although note that the irrationals aren't open either, though they are a countable intersection of open dense sets (viz. $\cap_{q \in \mathbf{Q}} (\mathbf{R} - q)$, where $\mathbf{Q}$ is the rational numbers)).

Actually, it does much more. For one thing, it allows us to completely neglect any functions which aren't nice: Suppose we decide that functions having a nice property are open dense. Then we decide that the same is true of some other nice property. We'd like to have both properties, of course, which cannot be guaranteed on the basis of only density. But the intersection of a finite number of open dense sets is open dense. We

don't use probability because there is no natural measure for $C^r(M,N)$, and no natural probability distribution. We would like to say that in a measure space of total measure 1, an open dense subset is also of measure 1, but strangely enough, even though open dense sets are very dense indeed, this does not have to be so. For example, one can remove a Cantor set of positive measure from the unit interval, leaving an open dense set of measure less than 1.

Also genericity is related to stability. There are numerous definitions of stability; we are concerned with *structural stability.* A function $f \in C^r(M,N)$ is *structurally stable* with respect to some equivalence relation (e.g. topological equivalence) if all sufficiently small perturbations of f (relative to the $C^r(M,N)$ topology) result in an equivalent function. In other words, f is not a freak, destroyed by the least perturbation. With respect to the equivalence determined by possessing the generic property, the openness of generic sets makes a function with a generic property structurally stable. In other words, small perturbations of f do not affect the presence of the generic property. And with respect to some other equivalence, the density guarantees that a structurally stable function will be equivalent to a generic one. Usually, structural stability is defined with respect to some topological equivalence relation. E.g., we can define 2 functions $f, g : M \to N$ to be *topologically equivalent* if there is a homeomorphism $h : M \to M$ such that $f = gh$. This is the situation of Fig. GM. Notice that topological equivalence guarantees that topological properties will be shared, e.g. h takes level sets of f to those of g, so the structural stability (with respect to this topological equivalence) of f would guarantee that the level set structure topologically remains unchanged under small perturbations. Now the perturbation could also be derived from motions of the observer, and we might be interested in some other feature, e.g. a derived boundary. For the purpose of analyzing the picture, we would probably want to focus attention on boundaries whose topological structure didn't change with small changes in viewpoint, or in the picture (e.g. noise), hence we would want to focus attention on generic pictures

and structurally stable features. Of course, we are interested in more than just topology in analyzing a picture, so that is not *all* we would have to consider, but it is a first cut at separating wheat from chaff.

Multiple color dimensions: the cases $n \geq 2$

We now carry on with the proof of the 2-color theorem for higher dimensions

Let $f : M^m \to M^n$, be C^r and regular at p. The analysis is based on the fact that at a regular point, if there is enough room in the range space, f is a diffeomorphism from a neighborhood U of p to $f(U)$. This is yet another version of the implicit function theorem. The idea of enough room can be made precise simply by requiring the Jacobian to be 1-1. This is the case for a regular point if the dimension of the range space is at least that of the domain space, i.e. if $m \leq n$, which is the situation for us if there are at least 2 color dimensions.

As before, the possible maps g_π which solve the matching problem are exactly those which take level sets to level sets. Since the g_π are diffeomorphisms, we can just study the maps of the level sets of, say, F_1, since they are equivalent by a given g_π to the set of all g_π. (To see this, consider Fig. (equiv). Let h be a diffeomorphism which takes level sets to level sets, i.e. which makes the diagram commutative, and define $g'_\pi = g_\pi \circ h$, so that any h gives us a g'_π. Likewise given such a g'_π, define $h = g_\pi^{-1} \circ g'_\pi$.)

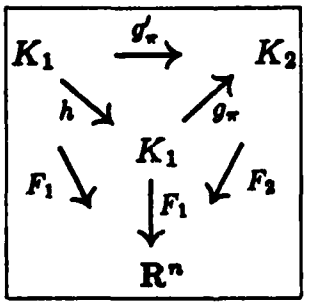


Fig. (equiv)

So, without any loss of generality, we can restrict our attention to the part of Fig. (equiv) shown in Fig. (equiv'), where we have replaced the notation K_1 by M^2 to help keep in mind that we are considering a 2-dimensional region:

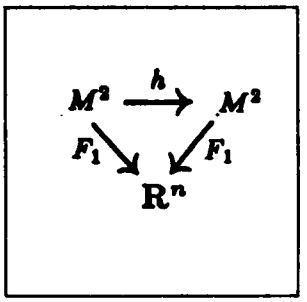


Fig. (equiv')

We pointed out earlier that any h which satisfies our conditions must take level sets to level sets. If F_1 is 1-1 for some point $q \in \mathbf{R}^n$, then the level set for that point is just a single point, and there is no choice in what h can do: following the lefthand F_1 arrow backwards, and likewise the righthand one, we see that h must take the single point $p = F_1^{-1}(q)$ to itself and no other. So the question of the uniqueness of h becomes one of studying how F_1 can fail to be 1-1.

First, let's look at how many points can be in $F_1^{-1}(p)$. By the implicit function theorem, since the dimension of the range (i.e. the color space) is at least that of the domain, the level set of a regular value is at most a discrete set of points. Since we are restricting ourselves to compact pictures, the level set must be a finite set (to avoid an accumulation point). Hence on a level set, g_π is constrained to be one of a finite number of permutations of the finite level set. Furthermore, since F_1 is a local diffeomorphism at a regular value, the permutation cannot jump around wildly among neighboring points, so that in fact g_π is a permutation of "sheets." I.e., let p be a regular value of F_1. Then $F_1^{-1}(p) = q_i, i \in \mathbf{Z}$. And there is some neighborhood U of p such that $F_1^{-1}(U) = V_i$, $q_i \in V_i$, and the V_i are disjoint. The V_i are then said to belong to different *sheets*, and the effect of g_π is to permute the V_i. It may happen that there is a path of regular points joining q_j to q_k, so that there is no global sheet as a set of points, although one can make an arbitrary partition (as is the case for the familiar integral power function in the complex plane). The sheets may be separated (and conceivably punctured) by critical points. Thus we are led to consider the topology of the critical sets, and the cardinality of the level sets.

Fortunately, others, notably Thom, Boardman, Mather, and Whitney were led to consider the same questions, beginning in the 1950's, and we now proceed to use some of their results. As it turns out, the higher dimensions are easier to deal with in our context, so we will start with them.

Regular points when $n \geq 3$

We are interested in studying how F_1 can fail to be 1-1. We know from the implicit function theorem that because $n \geq m$, F_1 is locally 1-1 at regular points. In other words, F_1 is a local embedding of its regular set into $\mathbf{R}^n$. But it may not be a *global* embedding, since the image may be self-intersecting. It is precisely at these self-intersection points that F_1 fails to be 1-1 on the regular set. For a regular p, $F_1^{-1}(p)$ consists of isolated points, so we can consider intersections of regular neighborhoods. What do these look like?

Theorem. (see e.g. [Hirsch 1976]) Let M, N be embedded submanifolds of $\mathbf{R}^n$. Then generically, $\dim M + \dim N - n = \dim M \cap N$, where a negative dimension means the intersection is empty.

We are interested in the case where M and N are the images in $\mathbf{R}^n$ of regular neighborhoods in M^m, so $\dim M = \dim N = 2$. From the above theorem we see that the intersection is generically of dimension $2, 1, 0$, and empty for $n = 2, 3, 4, 5$ resp. Thus if $n \geq 3$, the intersection set is generically of lower dimension in the embedded regular sets. h must be the identity other than on the intersection (since elsewhere F_1 is 1-1), and since removing a lower-dimensional subset leaves a dense set, h is generically 1-1 on a dense subset of the embedded regular sets. The continuity of h then guarantees unique continuation to the intersection, and there is again no choice in the behavior of h: it must be the identity. So for the regular points, we have disposed of all the cases of 3 or more color dimensions. Now we look at the critical sets, and *their* dimension.

The genericity of Morse functions can be generalized as follows.

Theorem(Critical set dimension). For an open dense subset of $C^\infty_g(M^m, M^n)$, the set of points of M^m where the Jacobian of f is of rank r

1) comprise a submanifold of M^m

2) $= \emptyset$ if $(m-r)(n-r) > m$

3) is of codimension $(m-r)(n-r)$ in M^m if $(m-r)(n-r) \leq m$

(X is of *codimension* k in Y if $\dim X + k = \dim Y$.)

Before we get involved in studying the critical sets for various color dimensions, we state 2 more closely related theorems which allow us to immediately understand the situations for 4 or more color dimensions. An immediate consequence of the critical set dimension theorem is the

Theorem (Whitney Immersion Theorem). If X, Y are smooth manifolds, with $\dim Y \geq 2 \cdot \dim X$, then maps with no critical points are open dense in $C^\infty(X, Y)$.

For a picture, $\dim X = 2$, so the above theorem applies when there are at least 4 color dimensions. In that case, it states that the typical picture won't have any critical points at all. Hence, typically there is only one "sheet" and no folds.

A further result is the

Theorem (Whitney 1-1 Immersion Theorem). If X, Y are smooth manifolds, with $\dim Y \geq 2 \cdot \dim X + 1$, then 1-1 maps with no critical points are residual (i.e. generic) in $C^\infty(X, Y)$.

So with at least 5 color dimensions, we can assume no color is used twice.

Returning to the critical set dimension theorem, in our case, $m = 2$, so what the theorem tells us is that the dimension of the critical set is respectively $1, 0$, and empty for $n = 2, 3, 4$.

By reasoning as we did for multiple points of the regular set, h, the diffeomorphism of Fig. (equiv') which leaves the picture invariant, has unique continuation to the critical set for $n \geq 3$, yielding the conclusion that h is generically unique when $n \geq 3$ (for $n = 2$ the 1-1 set need not be dense, so the conclusion wouldn't follow).

To summarize, we have thus far shown that h must be the identity for $n \geq 3$, and is at worst one of a discrete set of sheet permutations for $n = 2$. Now we will pursue the case $n = 2$ a bit further.

If we allow the support of a picture to be all of $\mathbf{R}^2$ or S^2, that is all we can say. (Consider, e.g., the function $z \mapsto z^k$ (for some $k \geq 2$) on the complex plane for the picture function. Then the sheets can be permuted leaving the picture invariant.) But a real picture must be finite in extent, so if we are considering subsets of the plane, a rectangle (i.e. a disc) is an appropriate domain to consider. If we are thinking about the sphere, then since we are restricting ourselves to occlusion-free regions, using the entire sphere would imply that there were no observable occlusions, which could only happen in the improbable events that only one object was illuminated, or that the observer could only see an object which completely enclosed him. Right now we are only concerned with the genericity of mappings of the plane, since we are in the context of general matching, so we will make no claims regarding the genericity of occlusion or illumination, though such an analysis is possible.

Let us now assume that the picture support we are considering is topologically a disc. Then h, being a homeomorphism, must map the boundary of the disc (a circle S) to itself. If f is 1-1 then h must be the identity. If not, then consider what must happen on this circle. h must be continuable along S, so for $p \in S$, $f^{-1}(p)$ must contain a constant number of points. This excludes the possibility of transverse crossings of $f(S)$. But transverse crossings for such a map are generic, so h must therefore generically be the identity. **QED** ∎

at does the 2-Color Theorem really mean?

ght seem that we have shown that monochrome stereoscopic vision is impossible. But y people with normal binocular vision have experienced stereoscopy with monochrome ires such as aerial surveys, stick figures of molecules, random-dot stereograms [Julesz , Julesz 1971], etc. What is more, there is evidence that color is not important uman stereopsis [Gregory 1977]. Machine stereo systems have been confined to ochrome pictures, and though they have not approached human performance, they been successful in extracting usable depth information.

it led to the 2-Color Theorem was the observation that a change in viewpoint leads complicated distortion of an object's picture. This distortion depends on surface e, viewpoint change, and imaging geometry and optics. The problem was to deduce distortion from the data, i.e. to solve the correspondence problem. What we studied the degree to which this problem can be solved purely from the topology, without idering the extra complexities of many possible geometric constraints. We confined attention to open sets free of singularities, i.e. areas without occlusions, which is of se only a part of the stereo vision problem.

were able to show that generically the monochrome problem is highly degenerate, we characterized the degeneracy. For the *color* problem, however, it turned out that ly topological considerations *were* enough to (generically) solve the problem, and that geometric information was therefore redundant. The conclusion is that there is a big rence between monochrome and color stereo; monochrome stereo requires and must fully incorporate geometric constraints to succeed in matching, while color stereo is ible without this, and can therefore use the imaging geometry in a different strategy.

have considered *generic* properties, so there can be infinitely many exceptions. But our results are for a generic subset of functions, they remain valid for small per- ations, and since generic sets are dense, every function can be approximated by a

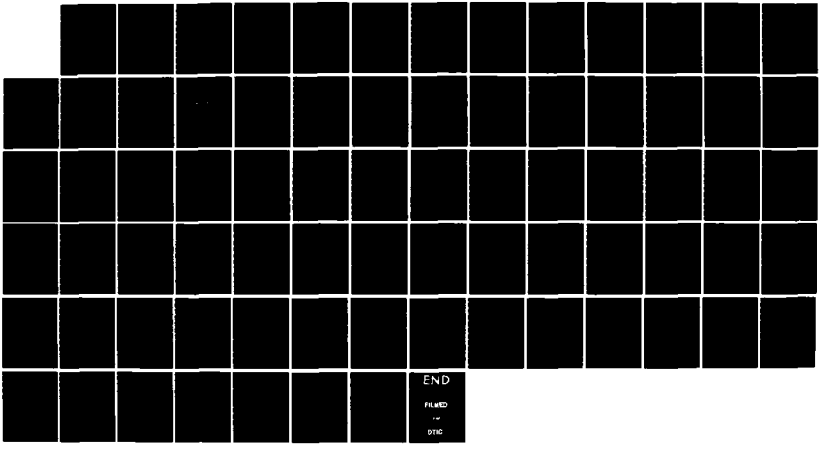
generic one to arbitrary precision. The results are about degeneracy, and the exceptions are invariably *more* degenerate. (This is simply because the exception sets are the inverse images of closed sets, e.g. places where a determinant is 0.) This means there are no special cases of monochrome pictures that are less degenerate. But it *is* possible to find *more* degenerate cases, so of course color pictures do not have to be uniquely matchable in special cases, the simplest of which is a region of constant color. Actual data contains noise and nonidealities, and digitization introduces degeneracy, so of course even with color one cannot expect perfect matching in a real world program, and surely one would still want to use constraints of imaging geometry to help the solution.

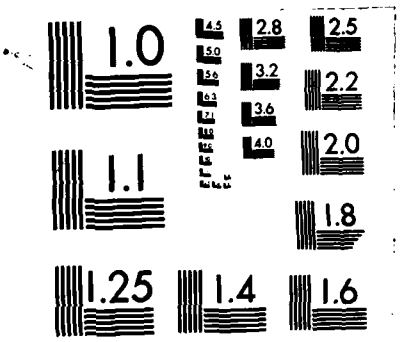
On the other side of the coin, a generic monochrome picture has isolated critical points, and a finite number of them for a bounded region. Since critical points must match critical points, finding this match is a combinatorial problem, which is made easier since the critical points have other attributes which are invariant, as we have discussed in the earlier section on differential topology for vision, and as we will discuss later in the section on topological invariants of the picture function. One can say essentially the same thing about level sets, *mutatis mutandi*. For a stereo pair of stick figures, then, most of the matching is between singular points associated with places such as branchings and terminations. The sticks themselves are individual level sets. Along those level sets, any stereopsis must come either from special knowledge of imaging geometry, i.e. the epipolar geometry relating the 2 retinas for a given state of convergence, interocular distance, focal length, eye rotation, retinal position, and focus, or additionally from gestaltist assumptions made by the visual system, such as an assumption of maximal simplicity. E.g., 2 horizontal black lines of different lengths presented one to each eye give no relative depth information about the region between endpoints, aside from continuity. However, this stimulus invariably gives a sensation of a straight line receding in space at some fixed angle. Similarly, a random-dot stereogram is assumed to be rectified, so the geometric constraint is known and easily used. There are a finite number of dots in each line, so the

1 is combinatorial, and there is no attempt to match the area within individual hich is completely degenerate. The degeneracy is ignored through an assumption licity in interpolation, i.e. it is assumed nothing new is happening within the dots, only true that nothing *knowable* is happening. More generally, we are concerned eas of maximal dimension taking values in a space of maximal dimension, in other in 2-dimensional patches that have a range of brightnesses or colors. The results almost entirely on these dimensions, so if we consider situations involving different ions, we must expect different results.

sons for machine stereopsis are, as we said above, that monochrome stereo must reful attention to correctly using geometric constraints, while the topology is le to find a match between level set structures of some appropriate measurement ce characteristic. On the other hand, color may offer a way to avoid this problem. s a large literature on machine stereo vision, and we will not attempt a review here. epresentative works are [Arnold, R.D. 1983], [Baker 1981], [Baker et al. 1983], d and Fischler 1982], [Gennery 1977, Gennery 1980], [Grimson 1980], [Hannah Ohta and Kanade 1983], [Marr 1982], [Marr and Poggio 1976, Marr and Poggio Moravec 1977, Moravec 1980], [Nevatia 1976], [Panton 1978], [Quam 1971]. For st part, the effects of geometry are not carefully considered; usually it is assumed ages are rectified, and no account is taken of possible distortion in the support ators used in the matching process. Using roughly vertical edges, i.e. places of rizontal gradient but small vertical gradient, renders some immunity, since these imally changed under the distortions of typical imaging situations. [Arnold, R.D. udied how the distribution of edge angles is related to geometry. [Baker 1981] did nline interpolation, but assumed rectified images; this is improved in [Baker et al. where epipolar geometry is explicitly considered. The epipolar geometry, however, rmined by a previous process of camera solving involving known interest point ondences. This permits epipolar line correspondence, but no correction is made

AD-A155 873 EDGE DETECTIÓN AND GEOMETRIC METHODS IN COMPUTER VISION 3/3
(U) STANFORD UNIV CA DEPT OF COMPUTER SCIENCE
A P BLICHER FEB 85 STAN-CS-85-1041 MDA903-80-C-0102
UNCLASSIFIED F/G 12/1 NL





for distortion of operater supports. [Panton 1978] made some use of epipolar constraints, and shaped the window, but doing this involved already having estimates of depth and surface shape. [Grimson 1980], following in the footsteps of [Marr and Poggio 1977], assumed rectified images, but found that this was not a reliable assumption and resorted to a vertical search to compensate for geometric factors. [Gennery 1977, Gennery 1980] was able to deal with imaging geometries, but did little about operator support distortion.

How does the constraint provided by epipolar geometry fit into the theory we have been developing? The epipolar constraint is quite analogous to the general matching constraint: 1-dimensional objects must be matched to corresponding 1-dimensional objects (we are confining ourselves now to monochrome pictures). For general matching the 1-dimensional objects are level sets. For epipolar matching, they are the epipolar lines. We do not study this in detail, but in the generic situation, one would expect these 2 families of curves to intersect each other transversely, and therefore give a discrete set of solutions for each point to be matched. It remains, however, to study what the degeneracies of this situation are. This is quite independent of the basic problem of determining the epipolar geometry. Generally this must be done by some combination of knowing the imaging parameters and solving a correspondence problem. Machine systems have relied heavily on the latter, so the problem is more subtle than may appear at first.

When is this analysis useful?

The F_1 and F_2 of Fig. (*') and the F_1 and F_2 of Fig. (GM), i.e. the "picture" functions we have considered in the general matching problem are assumed to be intrinsic to the object that is imaged. In practice, the absolute intensity levels or colors which one has available in a set of images are not completely precise, reliable, or consistent. E.g., they are likely to differ in bias (reference 0 level) and gain (measuring scale), suffer the effects of change in viewpoint and lighting on image irradiance, and contain digitization noise. Such considerations have discouraged people from using programs that try to match raw

measured intensity values directly, and instead have led to the use of derived values which are felt to be more stable.

Our results are *not* just statements about so-called *intensity matching*. The above theorems about matching are statements about intrinsic surface characteristics associated with points in images. They remain true even if the images themselves are not directly matchable; i.e. if our goal in matching is to match points that have the same value of an intrinsic function, then our theorems will govern the uniqueness of the match, regardless of how the actual images must be manipulated, or how they came about. If a derived function, truly intrinsic to the object, is to be matched, our results are just as applicable, providing, of course, that the new function is not computed in some degenerate way, destroying genericity (which would lead to even greater degeneracy in the solution). E.g., the digitization process cannot decrease the ambiguity, since it is a projection to a lower dimensional space. Since this mapping cannot be 1-1, it is unavoidable that the ambiguity will be increased, unless very special conditions occur. We have not studied the degradation imposed by digitization systematically here.

Extension to unknown bias and gain settings

What happens if we try to apply our analysis to functions which are *not* intrinsic to the surface? For certain kinds of ambiguity or lack of calibration, we would like to know that the data we get still allows the same uniqueness or degeneracy of match as with an intrinsic function. As an example of such a situation, we analyze what happens when gain and bias values are unknown. We have chosen this example because it is commonly believed that the uncontrollability of these parameters is a major impediment to intensity-based matching. We show that these extra degrees of freedom have no effect on the degeneracy or uniqueness of the matching problem. The extra ambiguity does, however, pose a greater challenge for a matching algorithm.

Before, we were concerned with the problem of Fig. (α): finding g_π such that $F_1 = F_2 \circ g_\pi$.

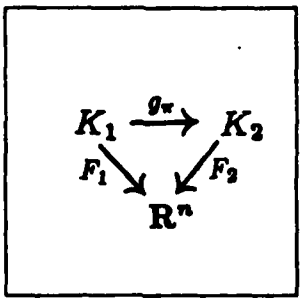


Fig. (α)

If a measurement of the variable x yields the value $ax + b$, then a is called the *gain* and b is called the *bias*.

Suppose that we observe the functions F_1 and F_2 as before, but now the bias and gain settings may be different between the observations, so we must first correct for the different settings before matching. This correction can be compressed into a single linear function, giving the new situation shown in Fig. (α-bias).

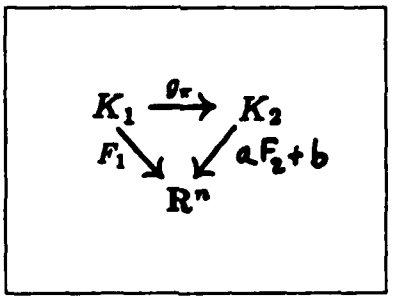


Fig. (α-bias)

The matching problem then becomes to find g_π such that $(aF_2 + b) \circ g_\pi = F_1$ for some $a \in \mathbf{R}, b \in \mathbf{R}^n$, and we are concerned with the question whether such a g_π is unique; i.e. whether there exists some other g_π which makes the diagram commute for perhaps some other values of a, b. Following the same reasoning as earlier, this is the same as asking for a diffeomorphism h which makes Fig. (equiv-bias) commute for some values of a, b, c, d.

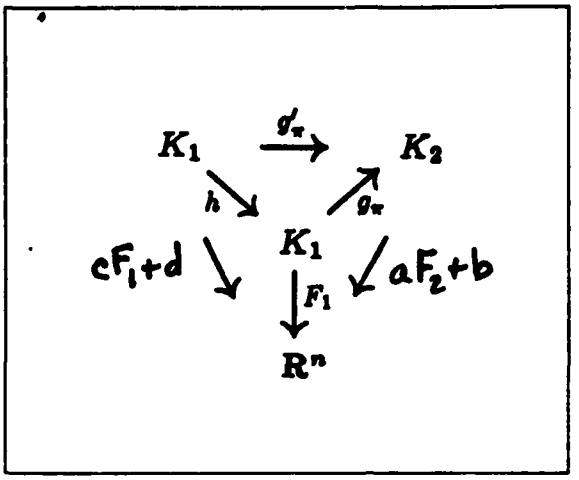


Fig. (equiv-bias)

We now will prove

Theorem. The conclusions of the 2-color theorem remain unchanged even for unknown bias and gain differences between pictures.

If we try to find h only for the situation that $c = 1$ and $d = 0$, we have exactly the problem we considered before, without gain or bias. So any h which satisfies our old conditions will also work if there is gain and bias error, although of course there may be even more h's for other values of c, d. Thus, the gain and bias matching problem is at least as degenerate as we have proved earlier for the "pure" problem.

The situation of Fig. (α-bias) can also be represented as

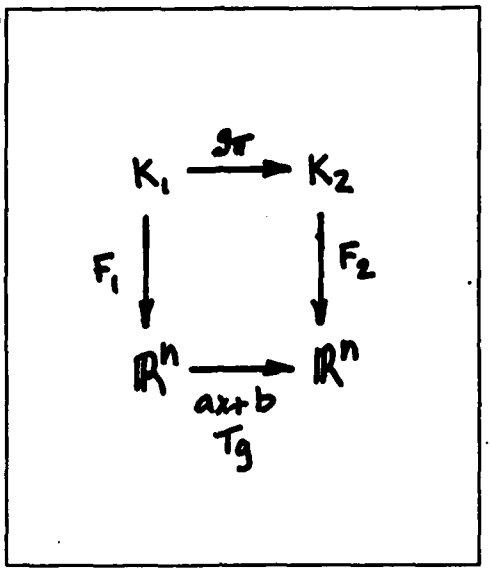


Fig. (α-bias')

where we have now included the unknown gain and bias parameters in a map

$$T_g : \mathbf{R}^n \to \mathbf{R}^n$$
$$y \mapsto ay + b$$

(Incidentally, one can take a to be some $n \times n$ matrix, to allow for different gains in different spectral bands, including linear crosstalk. In the absence of crosstalk, a is a diagonal matrix.) Then the analog of Fig. (equiv-bias) is Fig. (equiv-bias').

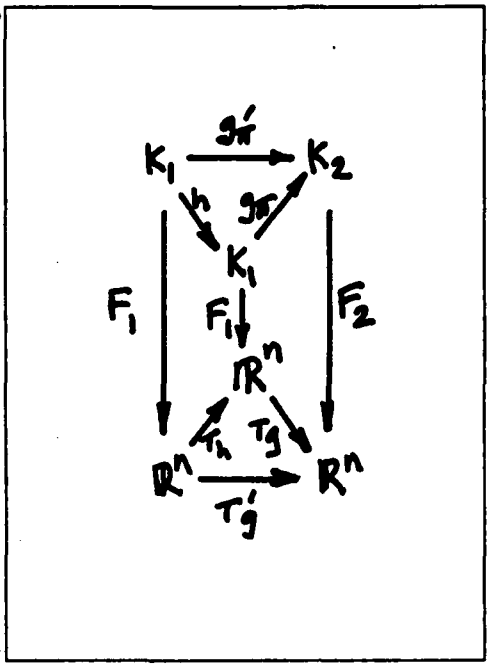


Fig. (equiv-bias')

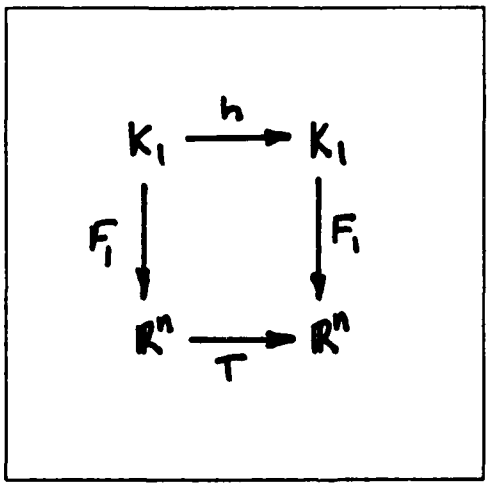


Fig. (equiv'-bias')

We see from this, as we earlier saw from Fig. (equiv) that the problem of uniqueness is equivalent to finding h such that the diagram in Fig. (equiv'-bias') commutes for some T.

When T is the identity, we have the problem which we analyzed earlier, so we are interested in what happens for nontrivial T. Using the same reasoning as before, we see that a necessary and sufficient condition for commutativity is that $\forall y \in \mathbf{R}^n \quad h(F_1^{-1}(y)) = F_1^{-1}(T(y))$ which we can write $h : F_1^{-1}(y) \mapsto F_1^{-1}(T(y))$. (Note that the inverse images are sets, not points, as F_1 may not be 1-1.) But there isn't any guarantee that $T(y) \in \text{Range}(F_1)$, even if $y \in \text{Range}(F_1)$! Since h is a diffeomorphism, whatever we say about h also goes for h^{-1}, so the first prerequisite for the existence of h is that $T(\text{Range}(F_1)) = \text{Range}(F_1)$. Suppose that $\text{Range}(F_1)$ is bounded. This must be so if K_1 is contained in a compact set; and in any case any real image would have a bounded range of values. Then it is easy to see that for scalar gain or no crosstalk, this cannot be the case for a nontrivial T. For the more general case when T has cross terms and/or $\text{Range}(F_1)$ is unbounded, it seems likely that T and F_1 would have to be very special, and hence not generic, for the range condition to hold. We illustrate the ready failure of the range condition for monochrome images in the following figure:

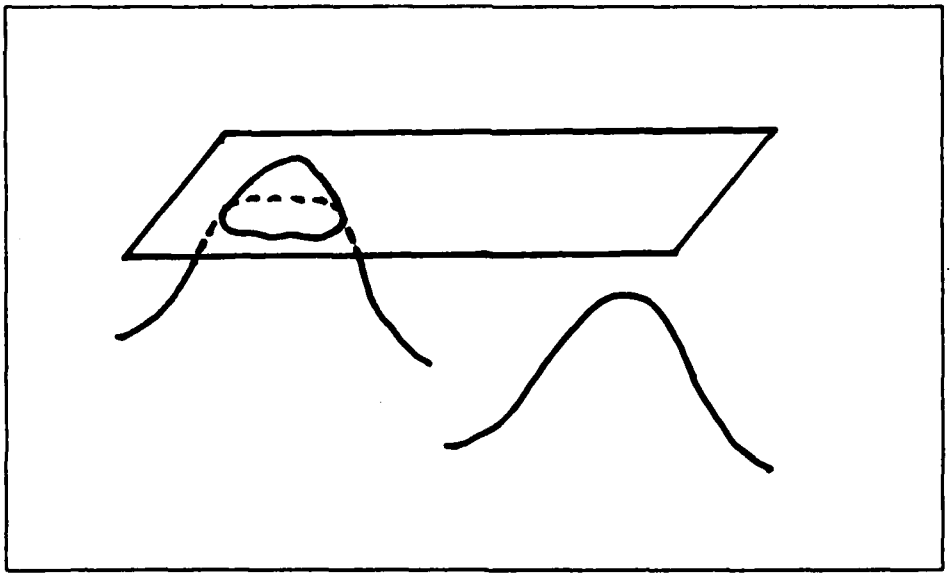


Fig. (Range condition)

The 2 humps represent a part of F_1, before and after the bias/gain change T. For the value represented by the plane slice, there may well be no corresponding value in the bias/gain-changed image.

Thus we have shown that even if we allow for the possibility that there may be unknown bias/gain changes between corresponding images, so that we are forced to do matching of values corrected for arbitrary bias/gain, our results remain unchanged. Furthermore, we have shown that for reasonable T's, T is unique; i.e., there is only 1 possible bias/gain transformation which allows matching. **QED** ∎

So far, though, we haven't addressed the question of discovering the correct T. Let's consider only the monochrome case, so that the bias/gain parameters a, b are both scalars, and assume that Range(F_1) is bounded. If the upper and lower bounds are known for both F_1, F_2, then it's clear that there is a unique T which takes corresponding bounds to each other (assuming $a \geq 0$, i.e. one image is not a negative of the other). Unfortunately, this would not work very well for real images, since noise and inconsistencies between images might result in meaningless end points for the ranges. Ideally, we would want to match topological features stably in the presence of noise, without the requirement for finding the bias/gain relation independently.

Referring to Fig. (α-bias'), we can state the matching problem in the presence of noise as follows. Looking for the ***best*** match means trying to find mappings g_π, T_g which optimize the value $\sigma(F_2 \circ g_\pi, T_g \circ F_1)$ of some similarity measure $\sigma : C^r(K_1) \times C^r(K_1) \to \mathbf{R}$. (Candidates: L^2 distance, cross-correlation, etc.) The measure should be chosen in such a way that the optimal g_π, T_g are in fact the most probable, given information about the statistics of the noise, the statistics of g_π and T_g, ***and the statistics of images***, i.e. some "probability" distribution on $C^r(K_1)$. The quotes are there because it is a difficult problem to define even a measure on an infinite-dimensional space; e.g. such spaces do not admit translation invariant measures. Also the optimization must be carried out over

the infinite-dimensional space of diffeomorphisms between K_1 and K_2. This suggests the use of a variational principle. Of course, simplifying assumptions can be (and are) made.

It's important to notice that the addition of noise does not change the applicability of our results. For if we are seeking the "true" g_π, i.e. the one which specifies the correspondence between the unadulterated images, then any equivalent g'_π that may exist as a consequence of our results will match the *identical* unadulterated images, hence will be just as good as g_π under any measure of the form of σ. Of course, the corruption of the noise may lead to further degeneracy. One expects that the similarity measure can be designed so that the matching process is stable. I.e., small amounts of noise should lead to small uncertainties in the match (modulo the topological ambiguities we have shown), and sufficiently small amounts of noise should not disturb topological properties of the solution.

Rather than tackle this difficult problem now, we sketch a possible (though simple-minded) way of finding T_g independently. In the presence of noise, some averaging method is called for. Instead of trying to match only extreme values [note that finding T is a matching problem in the range space], we might try to somehow match some kind of average ranging over all values. Fig. (Range condition) suggests one approach, which is reasonable if g_π is approximately an isometry (which is often the case in practice). The idea is that for each $y \in \text{Range}(F_1)$, the measure μ of $h : F_1^{-1}(y)$ should be the same as that of $F_2^{-1}(T_g(y))$. If we picture the slicing plane in Fig. (Range condition) as moving up and down, then what we are saying is that corresponding slices in the 2 images should have equal total arc length. We already know that generically, these slices will be 1-manifolds, so we are justified in using arc-length as our measure. Let $S_i(y)$ be the total arc length of $F_i^{-1}(y)$, for $i = 1, 2$. Then we can plot S_1 and S_2 as functions of the real variable y. Note that these are continuous, but will have discontinuities in derivative, corresponding to critical values of F_i. The graphs will look something like those in Fig. (Range).

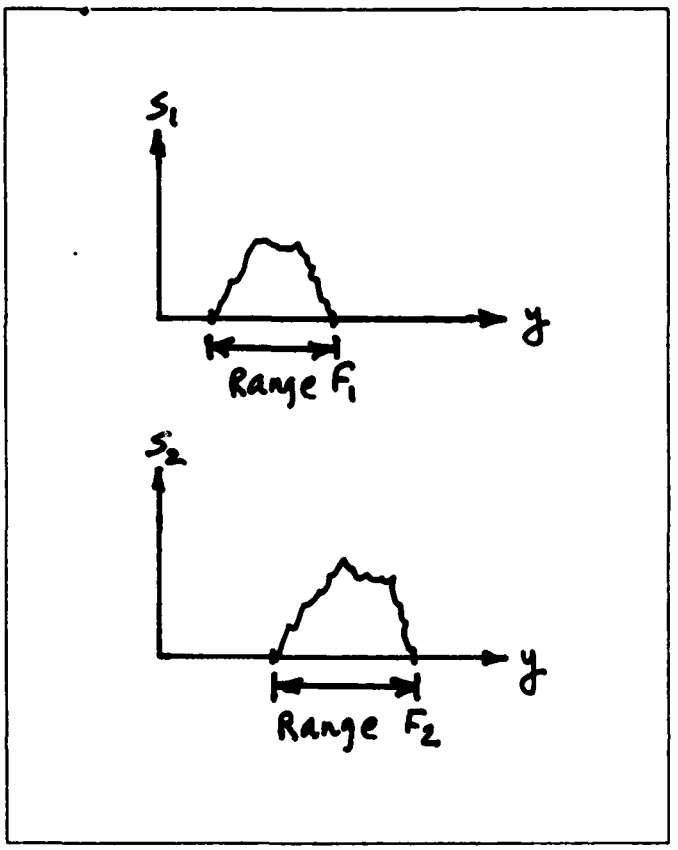


Fig. (Range)

If T is monotonic (which is of course the case for the 2-scalar bias/gain), then our problem is to match S_1 with S_2 by a linear map. In practice this would mean matching histograms of gray values. Since the search space is only 2-dimensional, this could be done by a brute force method. Alternatively, techniques exist for maximizing such matches, e.g. ***time-warping***. This technique, like histogramming methods in general, ignores topological and geometrical relationships.

Topological Invariants of the Picture Function

Introduction

It is well-known that computer scientists are fond of graph-theoretic approaches, so it is pleasing that the diffeomorphic invariants of (monochrome) pictures are well-represented as graph and tree structures. In this section, our goal is a representation of the picture topology to be used in the service of the matching problem. This requires us to present the applicable theory from differential topology, adapt it to our purposes, and allows us to make some observations along the way.

We review the definition of Smale diagrams, and establish the topological properties of level set trees. [Koenderink and van Doorn 1979] independently proposed level sets as topological invariants, but put them to different use, mainly as means to compute features of individual images, such as *metrons* and *aperture spectrum*. [Krakauer 1971] used a related structure for experiments in image analysis. He was interested in characterizing the shapes of the level sets at all image intensity values as a method of object classification. He used measures like eccentricity, region area, and scatter diagrams in an effort to identify various fruit, a fitting goal for a tree approach. He did not consider topological questions; the work was an attempt at direct interpretation from region-based descriptions, with little analysis of the nature of the image intensity function.

We, on the other hand, are concerned with the topology of the tree, its deformations and bifurcations as we move through the space of pictures, and the use of the bifurcations in handling noise. In general, we are concerned with using the level set tree as a representation, considering its generic behavior over the entire class of pictures, and using it in the matching problem. In a later section, we apply this structure to understanding *scale space* [Witkin 1983]. We also make some observations about the stability of zero-crossings. These applications are all new.

Smale diagrams and level set trees

We can formulate the general matching problem as a finite graph-matching problem. The nodes of the graph are to be the critical points of the image function F_1, and the (graph) edges are the gradient paths between the critical points. As we saw earlier, Morse theory tells us that generically the only critical points are maxima, minima, and saddles. More formally,

Definition. Let $f \in C^r(M^m, \mathbf{R})$ be a Morse function. For each critical point $p \in M^m$, define the ***stable*** and ***unstable manifolds***, resp. of p as follows.

$$W^+(p) = \{x \in M^m \mid \text{the uphill gradient line of } f \text{ leaving } x \text{ converges to } p\}$$
$$W^-(p) = \{x \in M^m \mid \text{the downhill gradient line of } f \text{ leaving } x \text{ converges to } p\}$$

Define a partial order $\leq$ (and hence a directed graph) among the critical points by $q \leq p$ if $W^+(p) \cap W^-(q) \neq \emptyset$.

Colloquially, $q \leq p$ means you can get from p to q by going downhill along gradients (the sense of the partial order is chosen to reflect $f(q) \leq f(p)$).

The *Smale diagram* [Smale 1967] of f is the ordered graph obtained by refining the preceding partial order so that $p \to q$ if $q \leq p$ and there is no r between p and q, i.e. such that $q \leq r \leq p$. A generalization of the same idea is known as a *Smale quiver* [Abraham and Marsden 1978].

spatial extent of the ignored features. These are the 2 parameters which are linearly combined in a linear smoothing operation, but here they are completely separable, and therefore accessible to reasoning machinery. Each subtree has its own characteristics, so a structure like this can be made as "adaptive" as you want; e.g. twiddles on very big humps might not mean too much, while the same twiddles on little humps might be quite important, though linear measures of local variation could be identical. This really has a philosophical basis in the principle of least commitment and in the AI paradigm of symbolic (thus nonlinear!) reasoning.

Not every bifurcation, and therefore not every smoothing, amounts to lopping off a subtree. As we saw before, we can also pass through a saddle connection. This implies that using the tree data structure requires some added sophistication, *viz.*, keeping track of where saddle nodes are relative to each other. More generally, there must be a notion of *measure of stability*—how far it is (in the function space) to a bifurcation.

We have described *all* the generic bifurcations of the level set structure. A generic scale space operator (i.e. a 1-parameter family of smoothers, whose $t = 0$ member is the identity) can therefore have only these bifurcations. A particular operator, however, might not have generic bifurcations; e.g. it might impose some special constraint that only allows special behavior. E.g., never creating zero-crossings is not a generic property for scale space smoothers (though it says nothing about the bifurcations of critical points). We have been able to show, though, that the generic critical point bifurcations of Gaussian scale space are not special, i.e. are the same as for generic perturbations, and are therefore among those we have described [Blicher and Omohundro 1984]. It remains to study which of these actually occur, and what are the unfoldings of the zero-crossing level set.

Now we can make some comparisons between Gaussian scale space and the level set topology tree. We have seen that zero-crossings are not stable, e.g. near a saddle in the picture. It is not clear whether the range of scales can fix this, for this depends on

the entire scale space is $\mathbf{R}^3$, and codimension 1 manifolds are surfaces. The locus of zero-crossings in scale space is one of the level sets of h.

To understand this level set of h is to understand zero-crossings in scale space. [Yuille and Poggio 1983] state that the Gaussian is the unique convolution kernel which does not create new zero-crossings with increasing t, under a number of regularity conditions. [Babaud, Witkin, Duda 1983] and [Hummel and Gidas 1984] also study this question.

We saw that the level set topology tree is a stable description of the level sets of f, and has simple, well-understood bifurcations. The nesting structure gives us an intrinsic, global criterion of relative scale, for if node x nests in node y, we know that x is a "twiddle" of y, and likewise for any sub-sub-nodes. Here's a way to think of this. Consider Fig. (topo) again. The saddle f is the 2nd type from the + column of Fig. (level). That means that we can take the stuff that sits above it on say the left side, cut it off at the f saddle level, and replace it by a simple cap. This could be done smoothly by some bifurcations (critical point annihilations) inside that side of the figure-8 of f. This is smoothing; highly nonlinear smoothing, however. Of course the exact single maximum cap that we get isn't uniquely defined, but why should it be? The picture doesn't have one cap or another; it has some complicated structure. The only justification for choosing a particular cap would be that it was somehow special. In the tree structure, what this amounts to is simply contracting a subtree to a single node. In a real picture, the tree structure is apt to be quite complicated, with great numbers of nodes. There will be an enormous number of ways to contract nodes to achieve grosser (smoother) representations. Actually, it may be better to consider the problem not one of multiple representations, but one of intelligent use of the single tree representation as a data structure. From that viewpoint, depth in the tree corresponds to degree of detail. However, that information must include more than nesting depth: there must be a measure of the significance of the ignored subtree. This can come from 2 things -- the size of the up and down excursions in the subtree (i.e. the range of leaf heights), and the size of the support of the subtree nodes, i.e. the

Scale Space

In our discussion of the works of Marr, Hildreth, Canny, and others, we saw that an important problem is the description of the picture at various scales. Small scales have the advantage of precision, and can pick out small features, but are susceptible to noise. Large scales can see large features that aren't visible in a small peephole, and can have good noise immunity thanks to averaging, but large linear operators confound space and intensity—they blur things.

[Koenderink and van Doorn 1979] proposed an *aperture spectrum* of an image be computed by convolving with a 1-parameter family of window functions. The aperture spectrum is the set of bifurcation values of the control parameter, in the usual parlance of bifurcation theory. [Crowley 1982, Crowley and Parker 1984, Crowley and Stern 1984] searched for some geometric features in data resulting from a sequence of convolutions with Gaussians, but did not consider geometric or topological theory. [Witkin 1983] also convolved with a 1-parameter family of Gaussians, and considered the bifurcations of zero-crossing topology in the combined control-behavior space (see [Poston and Stewart 1978]), i.e. the product space of the parameter and image, which he calls *scale space*. This is the usual approach of bifurcation theory, but [Witkin 1983] did not consider topological theory. The scale space operation is

$$h(t,x) = G_t(x) * f(x)$$

where f is the image, and G_t is a parametrized kernel. For scale space, a second derivative operation is required, so either f is the Laplacian of the image and G_t is a family of Gaussians, or f is the image and G_t is a family of Laplacians of Gaussians (x is a point in the picture space $\mathbf{R}^n$). Under these conditions, the object of interest is the locus of zeroes of h. When (t,x) is a regular point of h, the inverse function theorem tells us that $h^{-1}(0)$ is of codimension 1 near (t,x). This allows tying together zero-crossings at different scales, which was a major obstacle for many edge finders. For a picture on $\mathbf{R}^2$,

critical point) along the section of the tree on which it lies. When 2 nodes in the tree come together this way, giving 3 offspring of the combined node, we have a saddle connection. Also, buds can form anywhere (but generically away from other nodes), creating critical points, and leaf nodes can atrophy to nothing, annihilating critical points.

For a generic path like this, critical points can only be created or annihilated in pairs, for it is easy to construct arbitrarily small perturbations which separate the critical points into pairwise events. Recall that the Morse inequalities tell us that (with good behavior at the boundary), the sum

$$\sum_{k=0}^{n}(-1)^k\nu_k$$

must remain unchanged, so new critical points can only be created or annihilated in extremum-saddle pairs (the *saddle-node* bifurcation).

In the presence of noise, then, the level set structures of the 2 images may not be the same. They will differ by some sequence of the above bifurcations, which have simple effects on the level set tree. Equivalently, the 2 images will be connected by a path in function space which crosses some number of bifurcation frontiers. The matching problem can then be reduced to a minimal path or optimal tree matching problem (with labelled trees). The path to be minimized can be viewed as a sequence of level set topologies, equivalent to a sequence of bifurcations, or as the path in the function space itself. We have not studied the optimization criterion, but measures which could be taken into consideration are the number of bifurcations and the size of perturbations (in view of knowledge about the noise).

Occlusions result in localized but large differences between images. Globally, they could be handled by excision and pasting of tree parts (grafting?). This is delicate, however, since one must first study the global effects of excision and pasting. Another approach is to use a number of local analyses, for example for a number of regions selected by bump functions (which go from 1 to 0 smoothly to all orders in a finite space).

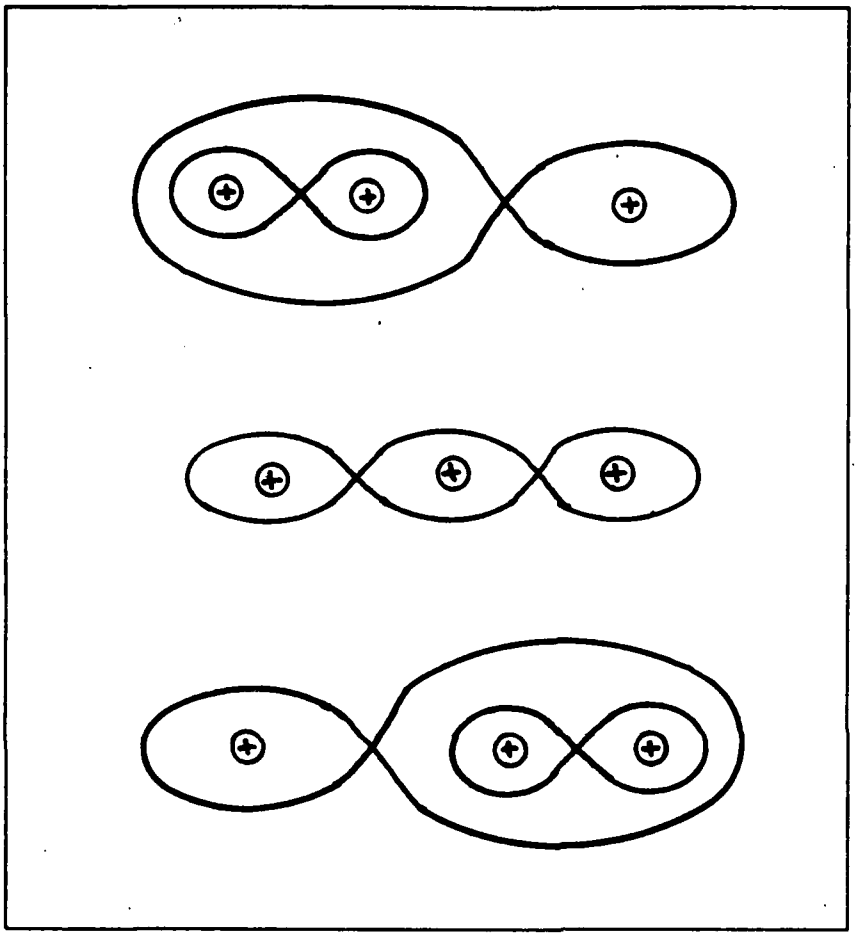


Fig. (saddle-connection)

The top part of Fig. (saddle-connection) shows a generic level set structure with 3 maxima. We can smoothly increase the height of the lower saddle until it is precisely at the same level as the other saddle, at which point the topology abruptly changes to the middle picture, called a ***saddle connection.*** As we continue raising the level of the saddle, we immediately get the bottom picture, which is again stable. A similar situation occurs when the other kind of saddle is involved.

If we think of the path through function space as a homotopy of maps to level set tree spaces, it is easy to visualize how the tree can change. Changing the height of a saddle corresponds to sliding a saddle node (a node in the tree, not the same as a saddle-node

talk about what a single noise signal can do to the topology; we are not going to attempt here a statistical study of these effects, so we will not consider, e.g., what the *average* effect on topology will be. It is, however, possible to do such a study; the ensemble properties must be considered over an appropriate space of control parameters, of course. E.g., these could be taken as parameters in the equation of a smooth surface. Examples of statistical yet topological studies of smooth functions are [Longuet-Higgins 1960], [Berry 1977], [Berry and Hannay 1977].

Think of a knob we can turn that gradually adds the smooth noise signal that has contributed to the function we are observing; this corresponds to a path in function space whose parameter is the amount the knob has been turned. Let $t = 0$ correspond to the unadulterated picture, and $t = 1$ to the picture with the noise we are actually observing. E.g., we could let $f_N(t) = f_0 + tN$, where f_0 is the unadulterated picture, N is the noise, and $f_N(t)$ is the adulterated picture at knob setting t. What happens to the level set topology as we turn our knob? Since it is stable, the topology changes at places where the function is not generic. These changes are called *bifurcations* or *catastrophes* and are completely classified by *singularity*, or so-called *catastrophe*, theory [e.g. Arnold 1984, Poston and Stewart 1978, Chow and Hale 1982, Iooss and Joseph 1980], providing simple rules which specify how the topology of f can (locally) change under such perturbations. For our case, 1-parameter families on $\mathbf{R}^2$, the situation is especially simple. Generically, there are only 2 ways this process can change the level set topology:

- Passing through a *saddle-connection*.

- Creating or annihilating critical points.

zero-crossing contours. Of course, one wants ultimately to segment the image, and the level set topology does not do that. We have argued that reliable segmentation can only be done after first getting a qualitative global understanding of the picture function, the type of understanding in which the level set topology is one element, and a beginning one. We wouldn't advise, therefore, to attempt segmentation at this stage. Nevertheless, if one insists that it is *approximately* the zero values one is interested in (we do not necessarily contend they *are* the correct thing to be interested in), then if one knows which critical points have values near 0, there are then a finite number of zero-crossing topologies depending on whether any given critical point has a positive or negative value. One could then use a constraint propagation procedure based on other information to select a particularly interesting subset of topologies.

Noise and bifurcations

Let's come back to the problem of noise changing the level set topology or the Smale diagram.

What kind of a model are we going to use for noise? We are working in the domain of smooth functions, so we are going to take any noise signal to be a smooth function, in keeping with the premise that the image irradiance is a smooth function. The statistical analysis of noise involves computing integrals, so a natural setting for statistics is in L^2, a space which contains mainly non-smooth functions. There are several reasons why we are justified in nonetheless taking our noise signal to be smooth. While it is convenient to do integration in L^2, physical signals are in reality bandlimited. Any imaging situation is well-modelled by a process that includes convolution with a smooth kernel, e.g. a Gaussian, i.e. it is impossible to avoid some amount of blurring. As we stated in detail earlier in the section Edge Localization in Both θ and x of the chapter Contributions to Edge Detection, a standard theorem [Lang 1969] tells us that the result of such a convolution is as smooth as the kernel, even if the signal is only in L^1. We are going to

is evident in zero-crossing edge finders when the connectivity shown in Fig. (Conn-a) is found in one image, while the connectivity of the corresponding region in the other image is found to be as in Fig. (Conn-b).

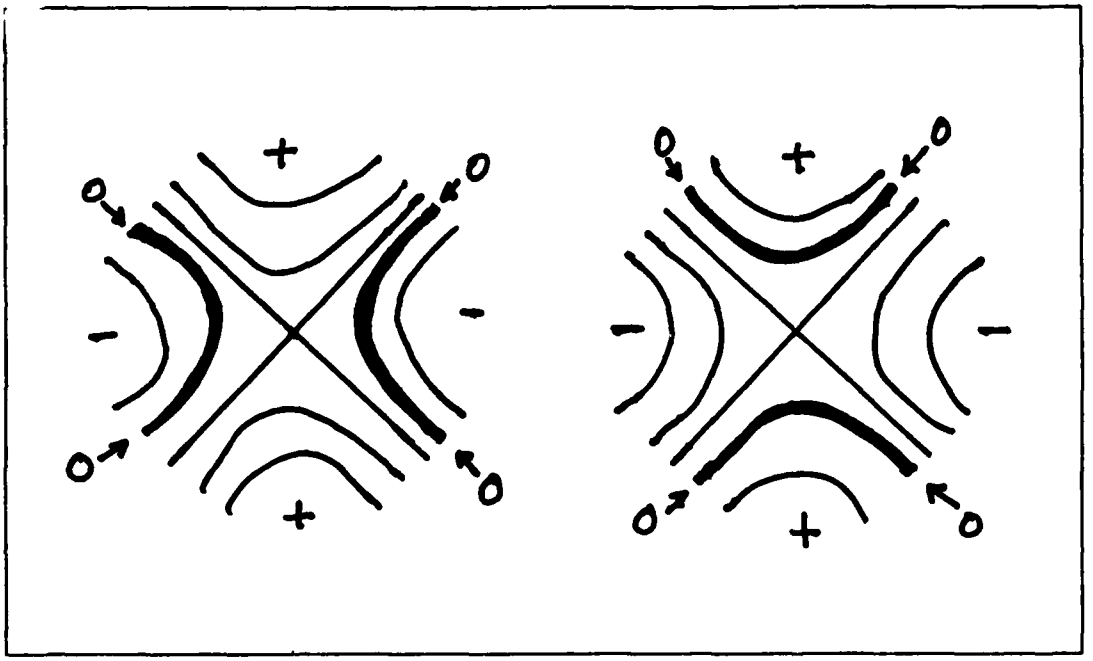


Fig. (Conn)

Referring back to Fig. (saddle), it is easy to see how this can happen with only a small amount of noise (or inexact mask-region correspondence between the images in a convolution). Suppose that the function whose zero crossings one is seeking looks like a saddle, with the critical value near 0. Then it is easy to imagine that in one image the zero plane would slice the saddle a little below the critical value, while if the other image has slightly smaller values, the slice would be above, yielding the grossly different (i.e. topologically different) connectivity patterns.

We stress that in this case, even though the zero-crossing connectivity is unstable, *the level set topology and the Smale diagram are unchanged.* Topologically, at least, the level set tree and the Smale diagram are more robust representations of the function than the

nesting diagram only classifies the coarser space of level sets. For example, in Fig. (topo), the 2 parts of the figure-8 of saddle d can be essentially interchanged by a 1-parameter family of diffeomorphisms, *viz.*, by shearing in a neighborhood of some level set just below d, just as in the proof of the 2-color theorem. In other words, choose a regular (circular) level set between d and f, grab the stuff above it, and rotate that stuff 180°, sliding along the level set. That can be made smooth by shearing in a neighborhood and splicing with a bump function. This doesn't change the level set topology, but it does interchange the roles of the components of the figure-8 in the Smale diagram.

Stability

The Smale diagram and the level set topology are ***stable*** for generic functions (i.e. Morse functions), i.e. they do not change under small perturbations. That means they are good ways to characterize the Morse functions, since the space of such functions is then partitioned into open regions (the boundaries are non-generic). Notice that stability is a criterion for ***robustness***, in that it means that there is some latitude for error which leaves the description unchanged. Right now we are interested in the level set structure of f rather than the topology of ∇f, so we will only discuss the level set topology. Unfortunately, the stability above, by itself, is not quite good enough for a practical system. What "small perturbations" really means is that the diagram will not change if the perturbation is small enough (values of derivatives are included in the measure). The problem is that there is no guarantee that noise corrupting the images will *be* small enough. Furthermore, for any given size of nontrivial noise, one can find a Morse function with a critical point so delicate that the level set topology will be changed by the noise (though not by ϵ times the noise for some ϵ, thus adhering to the stability theorem). Thus, a node (extremum) and a saddle might be introduced.

Instability of zero-crossings

A simpler cousin of this effect, not even changing the level set topology or Smale diagram,

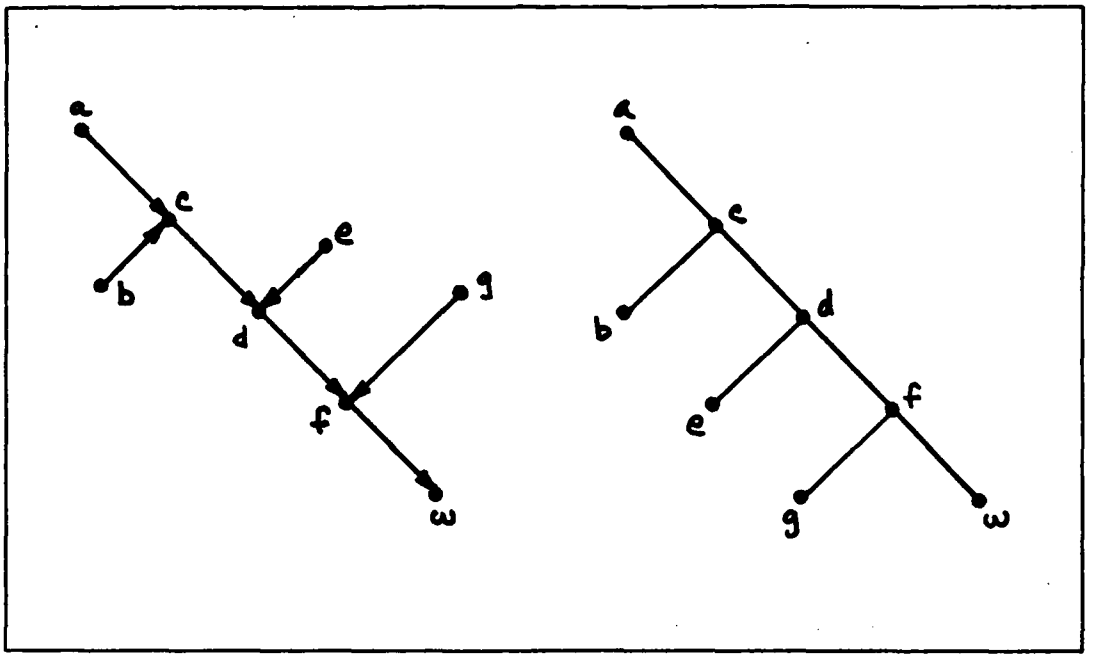


Fig. (tree)

Fig. (tree) shows 2 representations of the level set space for the image fragment in Fig. (topo). The tree on the left is drawn to show the nesting structure, relative heights, and extremum type of the critical points: the absolute height of each node in the tree is meant to correspond to the value of its corresponding critical point in the image, and the arrow is to be read "nests in." The right half of the figure shows just the bare tree, where a subnode nests in its parent node. Node a is the global maximum, and ω is the global minimum, in the following sense. For a compact manifold without boundary, a and ω are always critical points, but when there is a boundary, they correspond to the maximal and minimal closed level sets. This can be improved if the gradient is always transverse to the boundary, or smoothing with bumps can allow extension to the sphere.

The important difference between the level set tree and the Smale diagram for functions on 2-manifolds is this. Functions with the same level set topology can have different Smale diagrams, for the Smale diagram classifies the gradient flow of the function, while the

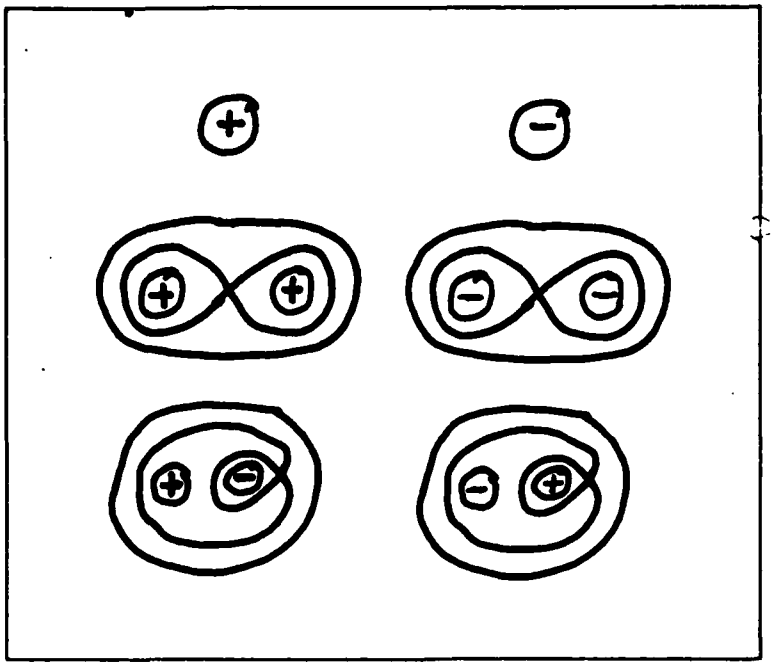


Fig. (level)

The generative rule is that any '+' can be replaced by anything from the '+' column, and similarly for '−'. The symbols represent a maximum and minimum, respectively. The 2 types of figure-8's are saddles (or more precisely, they are the separatrices associated with the saddle at the crossing). This leads to a representation of the topology of f as a tree, where the branch nodes are saddles and the leaves are extrema. In fact,

Lemma. As a topological space, the level set tree is homeomorphic to the space of level sets.

Proof. The topology is given by the local metric induced by the level values. It is easy to check that this is well-defined.

This structure is very similar to the Smale diagram, and the problem of matching 2 (monochrome) images is now equivalent to finding tree isomorphisms between the level set topologies (which preserve the image values at the nodes).

like drainage basins for lakes). The basins of attraction are separated by 1-dimensional boundaries (like the continental divide). Clearly every non-critical point must be in a basin or on such a boundary. If we turn things upside-down, the new basins of attraction are now *basins of repulsion* for the right-side up picture. [Nackman 1982, Nackman 1984] catalogs some of the behavior of functions on $\mathbf{R}^2$ which can be deduced from the partial order used to define the Smale diagram. (He calls this partial order the *critical point configuration graph*). Some examples of the modern mathematical approach to these features can be found in [Abraham and Shaw 1981, Gilmore 1981, Thom 1972, Hirsch and Smale 1974, Abraham and Marsden 1978, Smale 1967].

The Smale diagram is a diffeomorphic invariant of the vector field ∇f. The matching diffeomorphism g_π, however, carries with it the values of f, not of ∇f. And diffeomorphic equivalence of f and h is not enough to guarantee diffeomorphic equivalence of ∇f and ∇h, except for sufficiently small perturbations. If g_π is not too extreme, though, the problem of matching 2 (monochrome) images is equivalent to finding an isomorphism between the Smale diagrams of the images.

Instead of considering the topology of the Smale diagram, which classifies the gradient vector fields, we can consider the topology of the level sets of f. As [Koenderink and van Doorn 1979] have observed, these level sets observe simple rules in their nesting, which define a generative grammar. In fact, the only possible structures are shown in Fig. (level).

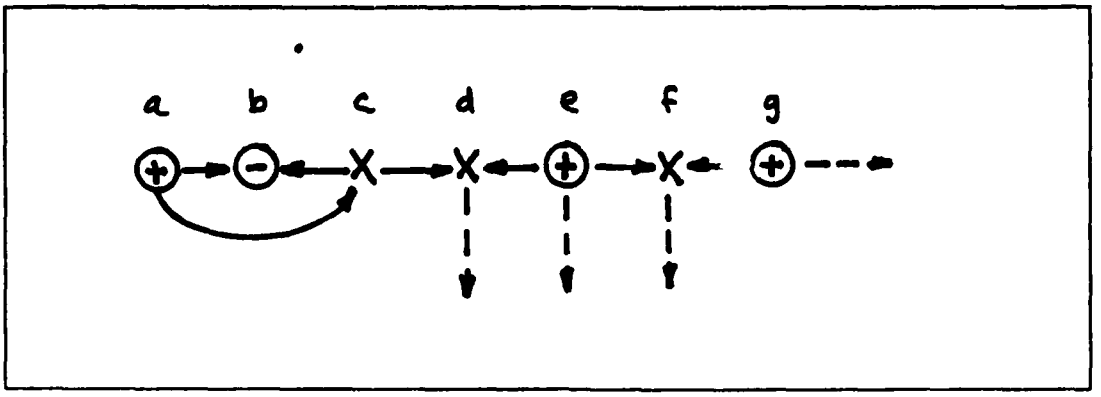


Fig. (partial order)

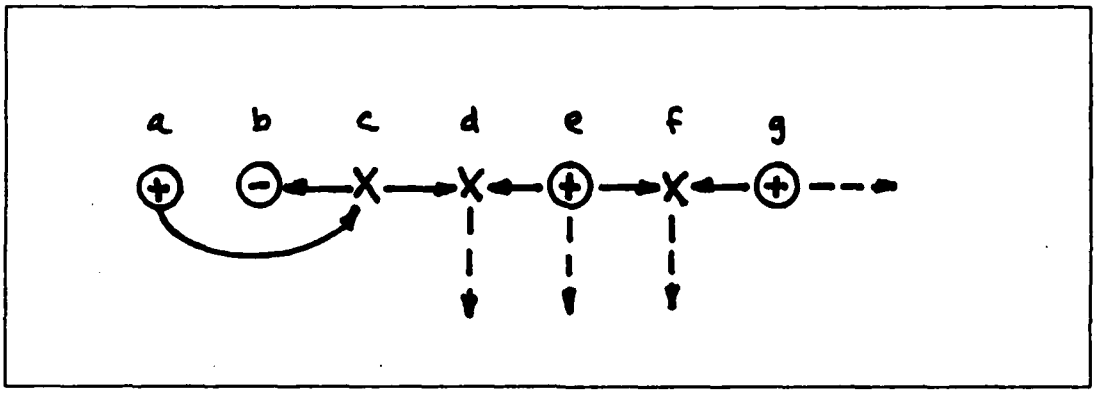


Fig. (Smale diag)

Fig. (topo) is an example of a level set structure one might find in an image. The partial order we have defined among its critical points is shown in Fig. (partial order), and the refinement to a Smale diagram is shown in Fig. (Smale diag). The dashed arrows represent partial order relationships which might exist with other critical points if we had extended the picture farther.

The entire topological structure of ∇f is given by its Smale diagram (possibly along with some orientation information) [Peixoto 1973]. If we know the Smale diagram, then we know how the critical points are connected, which lets us make deductions about the topology of the level sets between them. E.g., each critical point has a ***basin of attraction***, the set of all points whose gradients eventually lead to that critical point (just

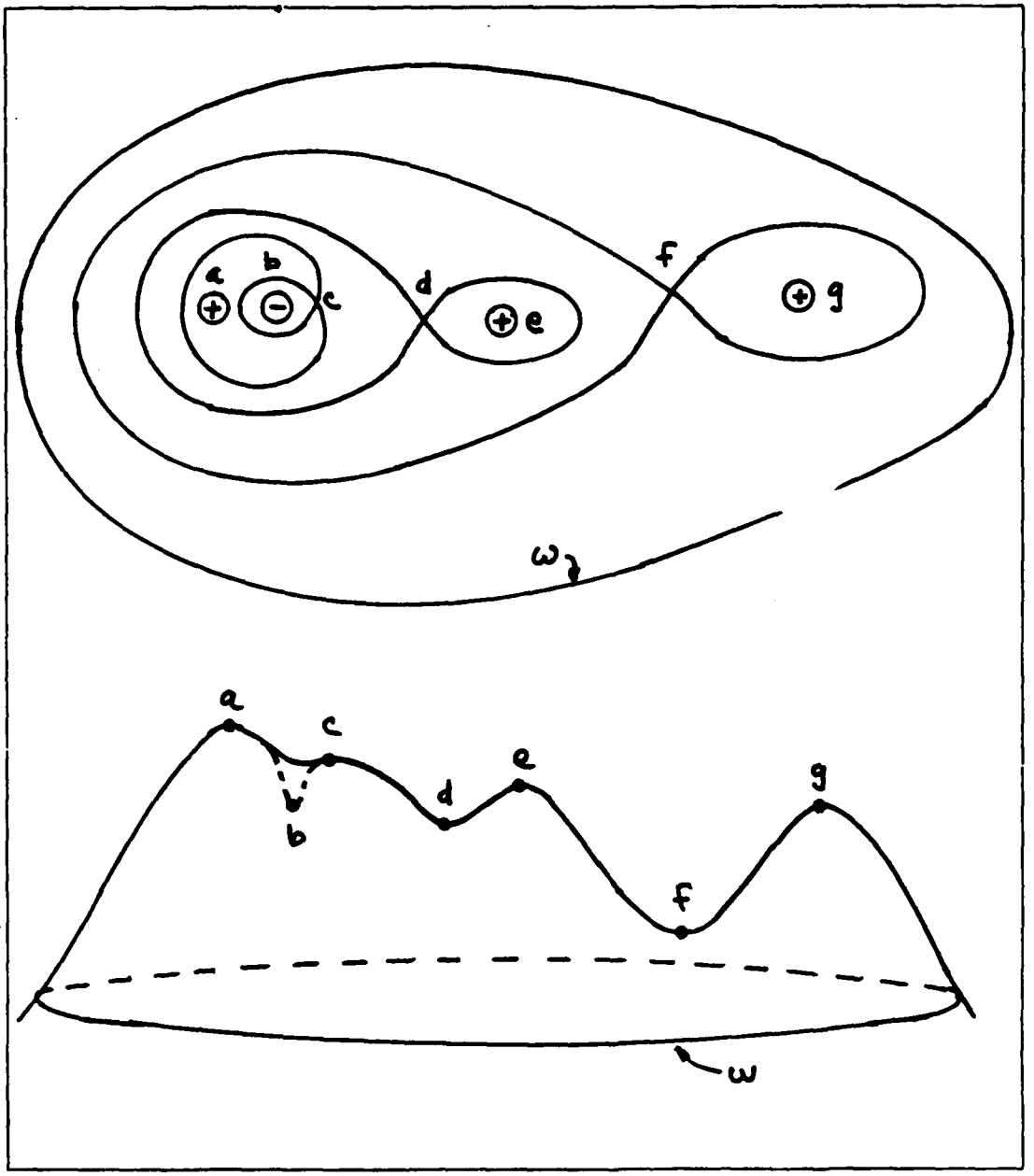


Fig. (topo)

the behavior of the smoothing near saddles. The level set topology, however, *is* stable. A case has been made that scale space allows tracking zero-crossings from coarser to finer resolution. The level set tree requires no tracking; the coarseness is established by depth in the tree (in the above sense) and the level sets at that depth are already precisely located. Gaussian scale space contains metrical information, for the result at a particular scale says something about extent. However, this metrical information is confounded with intensity information, as we have seen, so it is of limited value. The level set tree allows separating space and intensity. There is a double confounding of the metrical information, actually, because if we lift the Gaussian kernel to the surface that is projected to the picture, the nature of the kernel depends on the shape and orientation of the surface. This means that a change in viewer position, e.g., will give different results for the convolution, and while the zeroes of the Laplacian of the raw image are invariant, the zeroes of the smoothed version are not. The level set topology, of course, is invariant. Gaussian scale space is a particular class of bifurcations of the level set topology, a particular set of paths through function space, and so a specialization of the structure we are proposing. But the question is, why should the image values resulting from this particular smoothing be special?

Motion, Optic Flow, and Lie Algebras*

Introduction

For the past several years, many researchers have been investigating problems of moving objects and observers (see e.g., [Tsai and Huang 1984], [Prazdny 1981], [Prazdny 1983], [Buxton and Buxton 1983], [Nagel 1983], [Horn and Schunck 1980], [Tsai 1983a], [Tsai 1983b], [Prazdny 1980], [Bruss and Horn 1983], [Ullman 1979]). The paradigm of this research is based on the fact that a point moving in space projects to a point moving in the picture. The problem is then usually approached in 2 steps. First, to find the motion in the picture, the *optical flow*, you find corresponding points in 2 or more frames. Then, given this set of correspondences, either for a few or for many points, you solve some set of equations which yields the motion in space. These 2 subproblems have generally been approached separately; thus there are 2 classes of results: how to match points (correspondence), and how to compute motion from matches (e.g. how many corresponding points it takes). The correspondence problem, unfortunately, is subject to degeneracies, as we have shown above. E.g. at a single point, the image function and its time derivative tell us nothing about motion perpendicular to the gradient of the image function. If possible, then, it would be better to consider the problem as a whole, and avoid new difficulties created by a particular choice of subproblems, for, as we showed above, the correspondence problem is much harder without knowledge about the 3-dimensional changes that underlie the differences between pictures.

All the information which we have about the scene is contained in the time-varying picture, which is a function on some 2-dimensional space, as we said in more detail earlier. Our final goal is to deduce the shape, position, and motion of the 3-dimensional objects that give rise to this function. We want to approach this by looking *only* at the function itself, i.e. the time-varying image, and without the constraint that our intellectual path

*This work was done with the collaboration of Stephen M. Omohundro.

go by way of first finding some point motions in the plane.

The situation is this. Some rigid object is moving in space. Our imaging of it gives us a function, the image intensity function, which undergoes continuous distortions most everywhere. These distortions are a result of the motion and of the shape and position of the surface. The problem is to separate and quantify the sources of what we see.

The whole time course of the image has a vast amount of information in it, so it is easier to consider only parts of the information at once. One can look at what is happening over whole chunks of time, or only at a single instant. For differentiable situations, the differential theory is usually the easier, transforming nonlinear problems into linear ones, so that is where we start.

We prove some new theorems establishing how much picture information is necessary and sufficient to specify object motion. An important feature is that we do not assume that we can track individual points in the image, nor that we are given any of their velocities (i.e., the optic flow). The major result is the 6 point df/dt theorem, showing that generically* the values of df/dt at 6 points of the monochrome image f are necessary and sufficient to specify the motion of a given object. If we add color, we find that for 2 or more color dimensions, df/dt need only be known at 3 non-collinear points. Also, for 2 or more color dimensions, the optic flow is generically uniquely specified, in contrast to the monochrome case, where there is a 1-dimensional degeneracy.

We are going to use the notation of modern abstract geometry: Lie groups, Lie algebras, tangent planes, vector fields and bundles, etc. This lets us say things very compactly and simply, once the definitions are understood. Everything could also have been done without these abstractions (except maybe the use of genericity), solely in the language of classical calculus: vectors, rotation matrices, coordinate systems, etc, just as any

---

Recall from our discussion of the 2-Color Theorem earlier in this chapter, that a *generic* property is one which is true for a typical element of a space, i.e for a very dense subset of the space. For this section, we take this to mean an open dense subset.

computer program can be written in machine language, using absolute addresses. It is easier, though, to understand one written in a more abstract notation, especially if you don't happen to be its author. Instead of a maze of calculations, the reader is presented with simple (but rigorous) descriptions. Abstract mathematical treatment actually does more—it lets you understand a whole class of problems at once. Incidentally, this is really more than just analogy; the process of specifying concrete objects for abstractions can be automated into a compilation, so abstract notation can actually be *used* as a high-level programming language.

The mathematical structure

The situation is again that of Fig. (*'), except now the nature of the transformation g will be paramount.

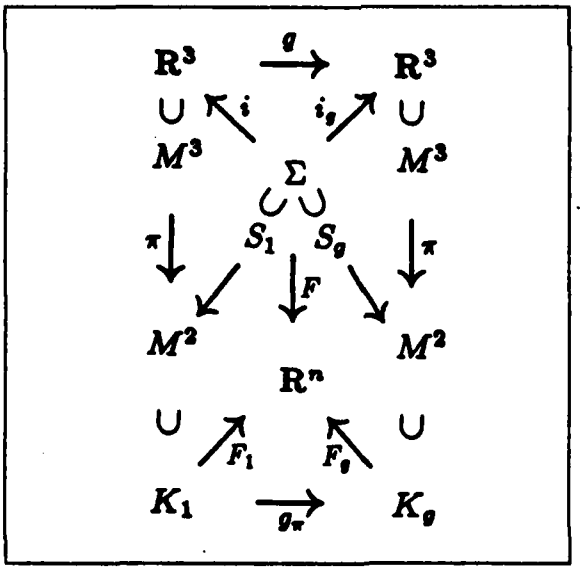


Fig. (*')

We are interested in rigid motions in $\mathbf{R}^3$, so $g \in E(3)$. The time evolution of the motion is then given by

$$\gamma : \mathbf{R} \to E(3)$$

i.e., as a path in the transformation group. In fact, γ defines a 1-parameter family of transformations. Since we are interested only in small changes from the current state, we take $\gamma(0) = I$, the identity in $E(3)$ (we could have done this anyway by using the group structure to translate back to the identity). For every t, γ gives a rigid motion of $\mathbf{R}^3$, since we are identifying $E(3)$ with the rigid motions of $\mathbf{R}^3$:

$$\gamma(t) : \mathbf{R}^3 \to \mathbf{R}^3$$

Each point of $\mathbf{R}^3$ is carried along with *this motion, and describes a path in* $\mathbf{R}^3$. In particular, every point of our surface of interest, embedded in $\mathbf{R}^3$, has such a path. Now apply the imaging projection, and restrict attention only to the visible surface of the embedded object. By composition, this leads to a path through each point that gets hit in the image. Now consider only a single time, $t = 0$. The structure we have presented thus far is summarized in Fig. (flow).

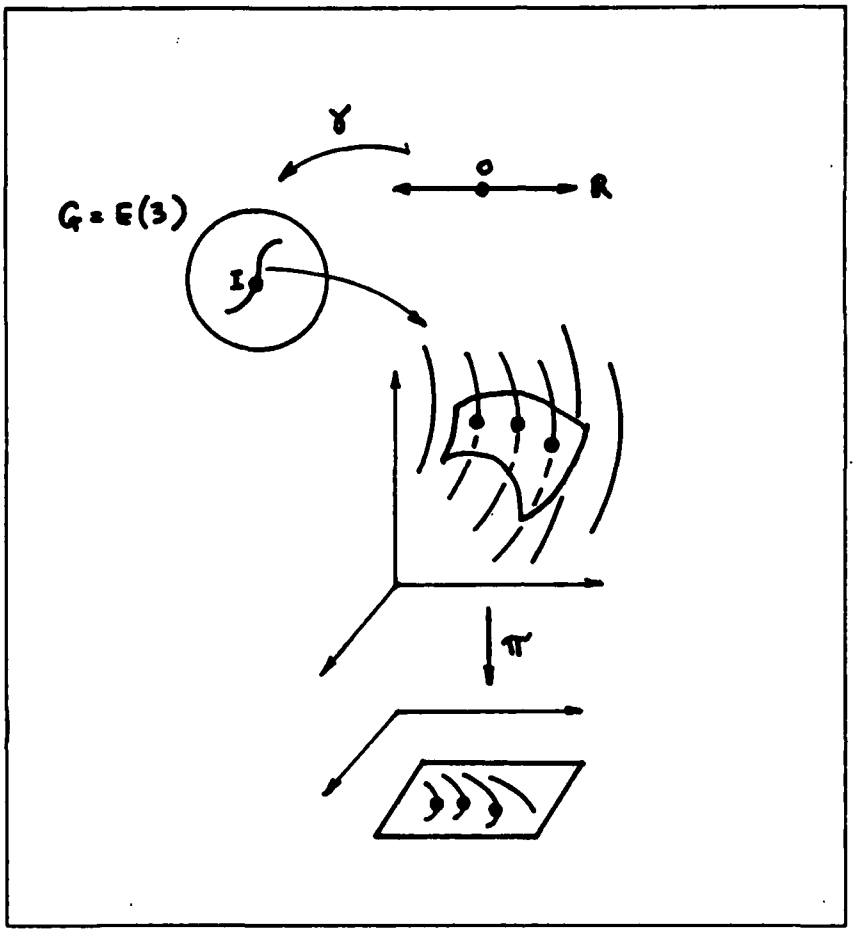


Fig. (flow)

Each such path in the picture has a velocity vector, and each point in the image has a path, so there is a vector field defined on the image. This is what is usually referred to as the *optic flow*, though it would be more consistent with mathematical terminology to call its integral, i.e. the paths in the image, the optic *flow*. We will reserve the term *optic flow* for this integral, i.e. the map $\varphi_t : U \to \mathbf{R}^2$ which specifies the paths of corresponding points in the picture with initial points in the region U, while using *optic velocity field* or *optic vector field* for its instantaneous velocities, the vectors $d\varphi_t/dt$. Similarly, the paths in $\mathbf{R}^3$ define a vector field on $\mathbf{R}^3$, and the path γ in $E(3)$ defines a tangent vector at the identity in $E(3)$.

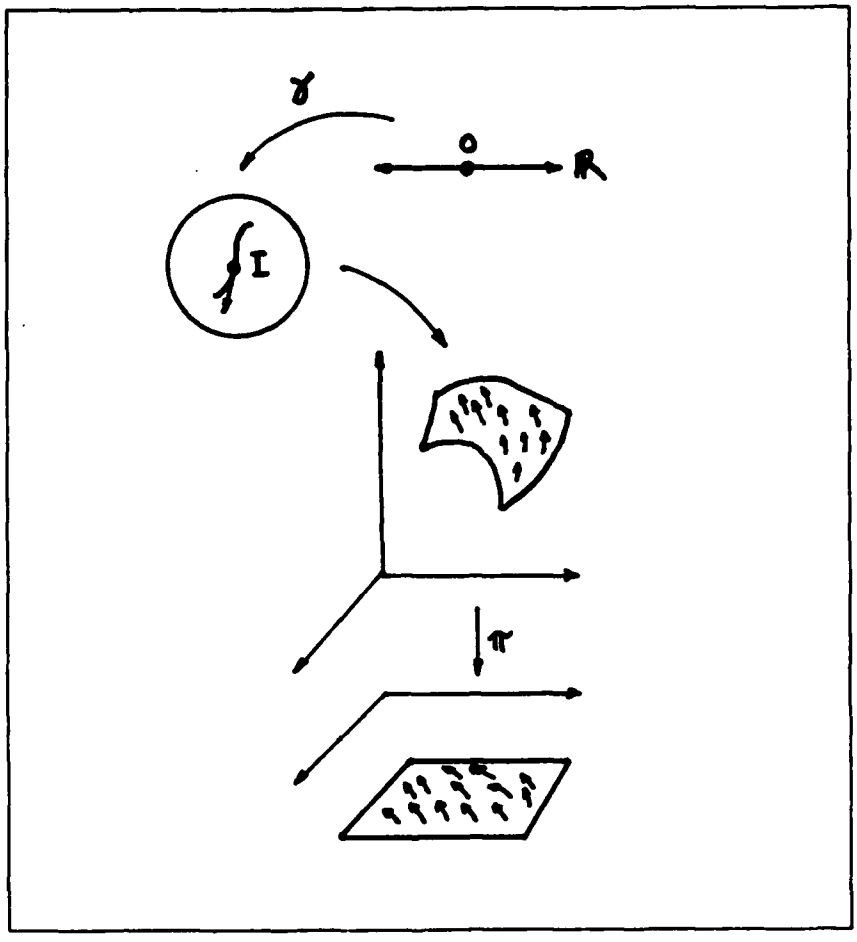


Fig. (vector fields)

The available data, however, is not the optical flow or vector field, but the time-varying picture function f_t which is just the projection of the intrinsic surface function F, under the same approximations we used in choosing a mathematical structure at the beginning of this chapter. Since we are considering only the differential theory, we regard our data as telling us only the instantaneous value f_0, and all the time derivatives at $t = 0$. This is the same as knowing the Taylor series for f_t. We will only use the 1st derivative for now. At a point p of the image, call the optic flow vector v. Then in a frame with velocity v at p in the image, f_t does not appear to change; the optic flow specifies the motion of corresponding points. Thus if we leave the frame fixed, we see that

$$\frac{d}{dt} f_t(p) = -D_v(f_t)(p),$$

where D_v means differentiation by the vector v, equivalent to $v \cdot \nabla$, so that

$$\frac{d}{dt} f_t(p) = -v \cdot \nabla f_t(p), \qquad (*)$$

(The more formal version of this theorem can be found on p. 91 of [Abraham and Marsden 1978], and was stated by Marius Sophus Lie in 1890. It is well-known in the context of optic flow; see e.g. [Horn and Schunck 1980, Ballard and Brown 1982].) Equation $(*)$ shows how it is that we only have partial information about v: we only know 1 component. We can immediately see, also, that if f had multiple dimensions, i.e. if there were more than 1 color dimension, we would have information about multiple components, and v would be uniquely determined for generic f. This is the differential version of the 2-color theorem we proved earlier. Finding optic flow, like matching, is much easier with color. We formalize this in

Theorem. (2-color theorem for optic flow) For a generic time-varying image function $f_t : M^2 \to \mathbf{R}^n$, the optic flow vector is uniquely specified at a generic point of the image if $n \geq 2$, i.e. for 2 or more color dimensions.

When we fix $t = 0$, each side of equation $(*)$ is just a number, so for each p we have a map

$$D_{\cdot}(f)(p) : v \mapsto \text{a real number}$$

We have thus defined a string of linear mappings (*v.f.* stands for *vector field*, *v.b.* for *vector bundle*):

tangent vector on $E(3) \mapsto$ v.b. section on object
$\mapsto$ v.f. on image $\mapsto$ vector at $p \mapsto$ real number

(We must consider *sections* of a vector *bundle* on the object rather than vector *fields* (sections of the tangent bundle) because the vectors we are interested in are tangent

vectors to paths in $\mathbf{R}^3$ going through points of the object. Since the paths generally do not lie in the object, their tangent vectors needn't be in the tangent space of the object, but rather are merely tangent vectors in $\mathbf{R}^3$.)

A *1-form* is a map which takes a vector field and spews out a scalar field, linearly at each point. I.e. at each point it linearly maps vectors to numbers. Thus it is dual to the notion of a vector field. A function f has a canonical 1-form, df associated with it by looking at how the function changes along paths. Consider a vector v at the point p. To define df at p, we must specify a number to which it will send v. v can be thought of as the tangent vector to some path, say $\gamma : \mathrm{I} \to M$, so that $v = \gamma'(0)$. Then we can define df by

$$df(v) = \frac{d}{dt}\Big|_{t=0} f\big(\gamma(t)\big)$$

df is sometimes called the *differential* of f. The space of all tangent vectors at a point is called the *tangent* space. A linear map from a vector space to the reals is called a *dual* vector, and the space of such maps, the *dual space* of the original vector space. The dual space of the tangent space is called the *cotangent* space, and its elements *covectors*. The disjoint union of the tangent spaces at all the points of a manifold is called the *tangent bundle*, and that of cotangent spaces, the *cotangent bundle*. The manifold that the vectors were originally tangent to is called the *base* space. Both bundles have natural structures as manifolds of dimension double that of the base space. A map which assigns to each point of the base manifold an element of its (co-)tangent space at that point, is called a *section* of the bundle. In the context of bundles, the *fiber* over a point is the (co-)tangent space of the original manifold at that point. A section chooses a point in each fiber. The tangent space at a point p of M is written T_pM, the tangent bundle TM, the cotangent space at p is T_p^*M, and the cotangent bundle T^*M. Thus df is a section of the cotangent bundle T^*M. ∇f, however, is a section of the tangent bundle, since it is vector-valued. It can only be defined if there is a canonical isomorphism between the tangent and cotangent bundles, e.g. if a metric is defined, or equivalently, a dot product. We will be confining our attention mainly to df. Instead of using tangent spaces to make a bundle, we can replace the role of the tangent space with an arbitrary vector space, yielding a *vector bundle*. A *Lie group* is a manifold which also has a group structure such that the group operation is a smooth map. Examples are matrices of nonzero determinant, and rotation groups. Like any other manifold, a Lie group has a tangent space at each point. Because of the group structure, though, vectors at the identity element of the Lie group can be moved around the manifold by the group action, so it is enough for most purposes to consider only the tangent space at the identity. This space is called the *Lie algebra* associated with the Lie group. It is an algebra because in addition to the vector space structure, there is a multiplication, called the *Lie bracket*. The bracket measures what the Lie group does to one vector as it moves it along in the direction specified by the other vector. The Lie algebra captures the infinitesimal behavior of its associated Lie group.

The Lie algebra $\mathfrak{g}$ of a Lie group G is a vector space which can be identified with the tangent space of G at the identity. $E(3)$ is a Lie group, and therefore associated with it is the Lie algebra $\mathfrak{e}(3)$; and since $E(3)$ is a 6-dimensional manifold, $\mathfrak{e}(3)$ is a 6-dimensional vector space. The tangent vector $\gamma'(0)$, which is the instantaneous motion, can therefore be thought of as an element of the Lie algebra $\mathfrak{e}(3)$.

We can do this for every path γ, hence for every element of $\mathfrak{e}(3)$, giving us a homomorphism from the Lie algebra $\mathfrak{e}(3)$ to sections of the vector bundle on the object, and likewise again to a Lie algebra of vector fields on the image of the object in the image plane. The

composition of these is a Lie algebra homomorphism. The sequence of linear maps can therefore be written

Lie algebra $\mathfrak{e}(3) \rightarrow$ v.b. sections on object
$\rightarrow$ v.f.'s on image $\rightarrow$ vectors at $p \rightarrow$ real numbers

This defines a map $\mathfrak{e}(3) \rightarrow \mathbf{R}$, i.e. an element of $\mathfrak{e}^*(3)$, the dual of $\mathfrak{e}(3)$.

Now we have enough machinery to attack some questions. The first question is whether there is enough information in df/dt to uniquely specify the instantaneous motion, for generic f. The instantaneous motion is an element of $\mathfrak{e}(3)$. As we just saw, for each point p of the image, the geometry defines an element of $\mathfrak{e}^*(3)$. The question then becomes whether we can span all of $\mathfrak{e}^*(3)$ by ranging over all points of the image, for knowing the value of applying a dual basis in $\mathfrak{e}^*(3)$ uniquely specifies the original vector in $\mathfrak{e}(3)$. $\mathfrak{e}^*(3)$ is 6-dimensional, so if this is possible, it is possible for 6 points corresponding to a dual basis. This doesn't say anything yet about finding the shape or position of the object; we only want to know whether we can recover the motion for fixed shape and position.

Theorem (6 point df/dt theorem). Let

$$\begin{aligned} f : \mathbf{I} \times U &\rightarrow \mathbf{R} \\ (t, p) &\mapsto f(t, p) \end{aligned}$$

be a time-varying picture for some time interval $\mathbf{I}$ around 0, and some neighborhood U in the image plane of regular values of the imaging projection of some 2-dimensional object embedded in $\mathbf{R}^3$. If f comes from the projection of a generic intrinsic function on an object undergoing rigid motion in $\mathbf{R}^3$, then the values of

$$\frac{\partial f}{\partial t}(0, p)$$

at 6 generic points $p \in U$ are necessary and sufficient to uniquely specify the instantaneous motion of the object.

Proof. We are in effect measuring the optic velocity field with our image function; this is what equation (*) says. To be able to tell the difference between different elements of $\mathfrak{e}(3)$, i.e. different motions, the mapping from $\mathfrak{e}(3)$ to velocity fields on the picture must be 1-1. Since the mapping is a vector space homomorphism, this is the same as saying it has no (nontrivial) kernel. The homomorphism

$$\mathfrak{e}(3) \to \text{v.b. sections on object}$$

has no kernel, because any kernel would leave the entire object fixed, but a rigid motion of $\mathbf{R}^3$ can leave at most a line fixed. So $\mathfrak{e}(3)$ is mapped 1-1 to sections of bundles on the object. Now we must show that the kernel of the homomorphism

$$\text{v.b. sections on object} \to \text{v.f.'s on image}$$

doesn't contain anything that comes from the previous map from $\mathfrak{e}(3)$. The kernel of the current map is just the sections whose vectors lie along the rays of projection to the picture. For orthogonal projection, vertical translation would of course be in this kernel, but we are assuming a projective projection, i.e. that the rays all meet at a point; for a planar retina this is the usual perspective projection.

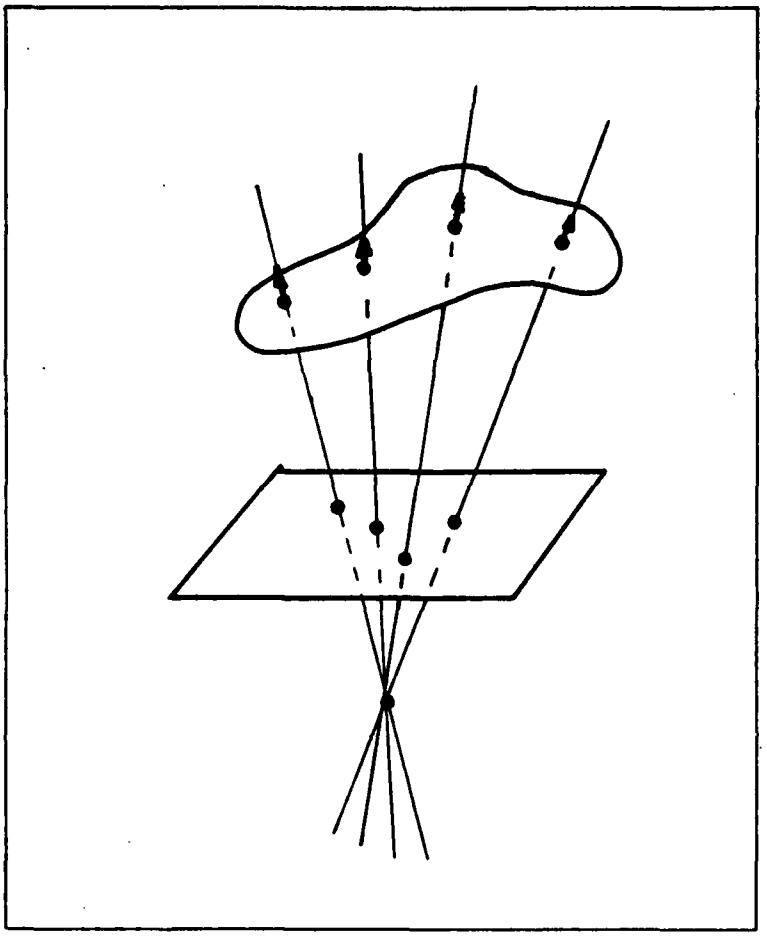


Fig. (kernel-rays)

We have to show that any such motion, where points move only along rays, cannot come from a rigid motion. This is easy to see; take 3 points on the object not all on the same line in $\mathbf{R}^3$:

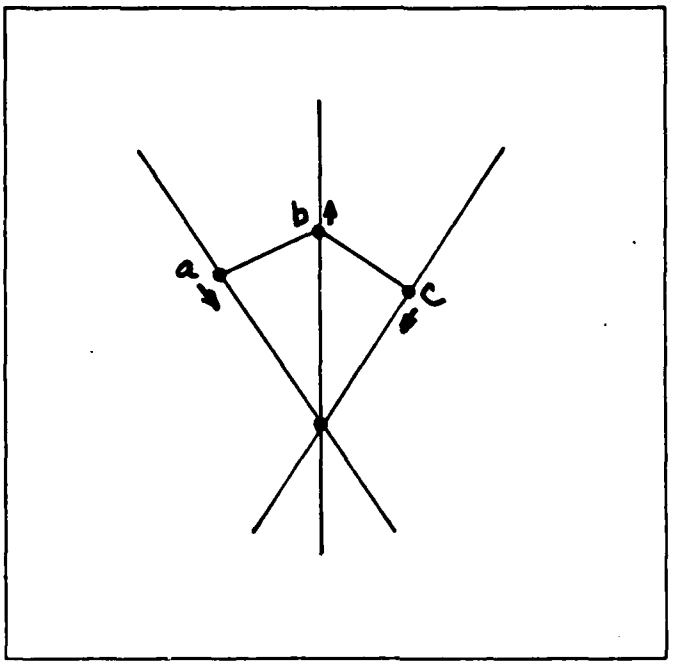


Fig. (3 points)

Since a rigid motion of $\mathbf{R}^3$ can only leave a single line axis (or nothing) fixed, at least 1 of the points must move, say a. If a moves down, b must move up, to keep their distance constant (rigid motion). Since b is moving up, c must move down. But then a and c are both moving down and therefore narrowing their distance, showing that the motion cannot be a rigid motion, i.e. the kernel of

$$\text{v.b. sections on object} \to \text{v.f.'s on image}$$

is not in the image of

$$\mathfrak{e}(3) \to \text{v.b. sections on object}$$

(except for 0, of course).

So we know that the composition

$$\mathfrak{e}(3) \to \text{v.f.'s on image}$$

has no kernel, i.e. is 1-1. This means that every rigid motion gives a unique optic velocity field, and the vector space of such fields is 6-dimensional.

Actually, we showed more than that. We showed that a generic set of 3 points cannot stay fixed in the image—we didn't even have to consider the whole vector field. The set of vectors at 3 such points in the image make up a 6-dimensional vector space, so what we showed is that the map

$$e(3) \to \text{vectors at 3 given points in image}$$

has no kernel, i.e. is 1-1.

That means that to specify a motion, i.e. an element of $e(3)$, we only have to figure out the optic velocity vectors at 3 points. A generic function, via equation $(*)$, tells us 1 component of each of the vectors (by genericity, the gradient is nonzero at all 3 points). If we had 2 generic functions, then we could recover both components of each of the 3 vectors by using equation $(*)$ for both functions (generically, the gradients will be linearly independent, i.e. in different directions at the 3 points). Parenthetically, we have just proved

Corollary (2 colors, 3 points). For generic f taking values in 2 or more color dimensions, the values of $\partial f/\partial t(0,p)$ at 3 noncollinear points $p \in U$ are necessary and sufficient to uniquely specify the instantaneous motion of the object.

Now we must show that 1 component at each of 6 points is as good as 2 components at each of 3 points.

We saw earlier that df defines an element of $e^*(3)$. Thus the geometry defines a map

$$T^*\mathbf{R}^2 \to e^*(3)$$

Duda and Hart 1973]

Duda, R.O. and P.E. Hart, **Pattern Classification and Scene Analysis**, Wiley, New York, 1973.

(cited on p. 14,15,19,41,81)

Fennema and Brice 1970]

Fennema, C.L. and C.R. Brice, "Scene analysis of pictures using regions", *Artificial Intelligence Journal* 1, 1970, 205-226.

(cited on p. 21,85)

Gennery 1977]

Gennery, Donald B., "A Stereo Vision System for an Autonomous Vehicle," *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, MIT, Cambridge, Massachusetts, August 1977, 576-582.

(cited on p. 178,179)

Gennery 1980]

Gennery, Donald B., "Modelling the Environment of an Exploring Vehicle by Means of Stereo Vision," Ph.D. thesis, Stanford Artificial Intelligence Laboratory, AIM 339, June 1980.

(cited on p. 154,178,179)

Gilmore 1981]

Gilmore, Robert **Catastrophe Theory for Scientists and Engineers**, Wiley-Interscience, New York 1981. [QA614.58.G54, ISBN 0-471-05064-4].

(cited on p. 193)

Golubitsky and Guillemin 1973]

Golubitsky, M. and V. Guillemin, **Stable Mappings and their Singularities**, (Graduate Texts in Mathematics 14), Springer, New York, 1973. [QA613.64.G64,

March 1955.
(cited on p. 9)

[do Carmo 1976]
do Carmo, M.P., **Differential Geometry of Curves and Surfaces**, Prentice-Hall, Englewood Cliffs, N.J., 1976.
(cited on p. 45)

[Dreschler and Nagel 1981a]
Dreschler, L. and H.-H. Nagel, "Volumetric Model and 3D-Trajectory of a Moving Car Derived from Monocular TV-Frame Sequences of a Street Scene," Report IfI-HH-M-90/81, Fachbereich Informatik, Universität Hamburg.
(cited on p. 18,48)

[Dreschler and Nagel 1981b]
Dreschler, L. and H.-H. Nagel, "Volumetric Model and 3D-Trajectory of a Moving Car Derived from Monocular TV-Frame Sequences of a Street Scene," *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* (IJCAI-81), August 1981, Vancouver.
(cited on p. 18,48)

[Duda and Hart 1971]
Duda, R.O. and P.E. Hart, "A Generalized Hough Transformation for Detecting Lines in Pictures," SRI AI Group Tech Note 36, 1971.
(cited on p. 19,81)

[Duda and Hart 1972]
Duda, R.O. and P.E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," *Comm. ACM* 15, no. 1, 1972, 11-15.
(cited on p. 19,81)

[Crowley 1982]

Crowley, James L., "A Representation for Visual Information," Ph.D. dissertation, Robotics Institute, Carnegie-Mellon University, 1982.

(cited on p. 202)

[Crowley and Parker 1984]

Crowley, James L. and Alice C. Parker, "A Representation for Shape Based on Peaks and Ridges in the Difference of Low Pass Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 2, March 1984, 156–170.

(cited on p. 202)

[Crowley and Stern 1984]

Crowley, James L. and Richard M. Stern, "Fast Computation of the Difference of Low Pass Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 2, March 1984, 212–222.

(cited on p. 202)

[Davis 1973]

Davis, L., "A Survey of Edge Detection Techniques", TR-273, Univ of Md, Computer Science Center, 1973.

(cited on p. 24,33)

[DeBoor 1978]

DeBoor, C., **A Practical Guide to Splines**, Springer, 1978 (Vol. 27 in Applied Mathematical Sciences series).

(cited on p. 54)

[Dineen 1955]

Dineen, G.P., "Programming Pattern Recognition," *Proc. WJCC*, 94–100,

(cited on p. 4).

[Bruss and Horn 1983]

Bruss, Anna R. and Berthold K.P. Horn, "Passive Navigation," *Computer Vision, Graphics, and Image Processing*, 21, 1983, 3–20.

(cited on p. 206)

[Buxton and Buxton 1983]

Buxton, B.F. and Hilary Buxton, "Monocular Depth Perception from Optical Flow by Space Time Signal Processing," *Proceedings of the Royal Society of London B*, vol. 218, 1983, 27–47.

(cited on p. 206)

[Canny 1983]

Canny, John Francis, "Finding Edges and Lines in Images," MIT Master's thesis, MIT AI-TR-720, June 1983.

(cited on p. 18,42,48,54,55,56,58,62,63,112)

[Chow and Kaneko 1972]

Chow, C.K. and T. Kaneko, "Boundary Detection of Radiographic Images by a Threshold Method," in **Frontiers of Pattern Recognition**, S.Watanabe, Ed., Academic Press, New York, 1972, 61–82.

(cited on p. 24)

[Chow and Hale 1982]

Chow, Shui-Nee and Jack K. Hale, **Methods of Bifurcation Theory**, vol. 251 of *Grundlehren der mathematischen Wissenschaften*, Springer-Verlag, New York, 1982. [QA372.C544, ISBN 0-387-90664-9].

(cited on p. 199)

J. Phys. A: Math. Gen., vol. 10, no. 11, 1977, 1809–1821.
(cited on p. 199)

[Binford 1970]
Binford, Thomas O., "The TOPOLOGIST," Internal Report MIT-AI, 1970.
(cited on p. 20)

[Binford 1981]
Binford, Thomas O., "Inferring Surfaces from Images," *Artificial Intelligence*, 17, 1981, 205–244.
(cited on p. 18,102,113,114,122)

[Blicher and Omohundro 1984]
Blicher, A. Peter and Stephen M. Omohundro, "Convolution with Gaussians has Generic Bifurcations," In preparation.
(cited on p. 80,204)

[Brice and Fennema 1970]
Brice, C.R. and C.L. Fennema, "Scene Analysis Using Regions," Artificial Intelligence Group Technical Note 17, Stanford Research Institute, April 1970.
(cited on p. 21,85)

[Bröcker and Lander 1975]
Bröcker, T. and L. Lander, **Differentiable Germs and Catastrophes**, London Mathematical Society Lecture Note Series 17, Cambridge University Press, Cambridge, 1975. [ISBN 0-521-20681-2].
(cited on p. 162)

[Brooks 1981]
Brooks, Rodney A., "Symbolic Reasoning Among 3-D Models and 2-D Images," *Artificial Intelligence*, 17, 1981, 285–348.

[Ballard and Brown 1982]

Ballard, Dana Harry and Christopher M. Brown, **Computer Vision**, Prentice-Hall, Englewood Cliffs, 1982. [TA1632.B34, ISBN 0-13-165316-4].

(cited on p. 212)

[Beaudet 1978]

Beaudet, P.R., "Rotationally Invariant Image Operators," in *Proceedings of the Fourth International Joint Conference on Pattern Recognition (IJCPR-78)*, (Kyoto, Japan, November 7 10, 1978), 579-583.

(cited on p. 18,19,44,47,48,49,50)

[Babaud, Witkin, Duda 1983]

Babaud,J., Andrew Witkin, and Richard Duda, "Uniqueness of the Gaussian Kernel for Scale-space Filtering," Fairchild Laboratory for Artificial Intelligence Research, Fairchild TR 645, FLAIR Memo 22, 1983.

(cited on p. 203)

[Barnard and Fischler 1982]

Barnard, S., and M. A. Fischler, "Computational Stereo," ACM Computing Surveys, vol. 14, no. 4, December 1982.

(cited on p. 178)

[Berry 1977]

Berry, M.V., "Focusing and Twinkling: Critical Exponents from Catastrophes in Non-Gaussian Random Short Waves," *J. Phys. A: Math. Gen.*, vol. 10, no. 12, 1977, 2061-2081.

(cited on p. 199)

[Berry and Hannay 1977]

Berry, M.V. and J.H. Hannay, "Umbilic Points on Gaussian Random Surfaces,"

[Arnold, V.I. 1983]

Arnold, Vladimir Igorevich, "Singularities of Systems of Rays," *Uspekhi Mat. Nauk*, 38:2, 1983, 77–147 Available in English translation as: *Russian Math. Surveys, 38:2, 1983, 87–176.*

(cited on p. 148,223)

[Arnold 1984]

Arnold, Vladimir Igorevich, **Catastrophe Theory**, translation of **Teoriia Katastrof**, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984. [QA614.58.A7613 1984, ISBN 0-387-12859-X (USA), ISBN 3-540-12859-X (FDR)].

(cited on p. 148,199)

[Baker 1981]

Baker, H. Harlyn, "Depth from Edge and Intensity Based Stereo," University of Illinois, Ph.D. thesis, September 1981; also available as AIM-347, Stanford Artificial Intelligence Laboratory, September 1982.

(cited on p. 154,178)

[Baker et al. 1983]

Baker, H. Harlyn, Thomas O. Binford, Jitendra Malik, and Jean-Frederic Meller, "Progress in Stereo Mapping," *Proceedings of the DARPA Image Understanding Workshop*, Arlington, VA., June 1983, 327–335.

(cited on p. 155,178)

[Ballard and Sklansky 1976]

Ballard, Dana Harry and J. Sklansky, "A Ladder-structured Decision Tree for Recognizing Tumors in Chest Radiographs," *IEEE Transactions on Computers*, vol. C-25, 5, May 1976, 503–513.

(cited on p. 19,83)

References

[Abdou 1978]

Abdou, I.E., "Quantitative Methods of Edge Detection," Ph.D. thesis, University of Southern California, July 1978. Also USCIPI Report 830.

(cited on p. 14,17,31,33,39,41)

[Abraham and Marsden 1978]

Abraham, Ralph H. and Jerrold E. Marsden, **Foundations of Mechanics**, 2nd ed., Benjamin/Cummings, Reading, Mass., 1978. [QA805.A2 1977, ISBN 0-8053-0102-X].

(cited on p. 160,163,190,193,212)

[Abraham and Shaw 1981]

Abraham, Ralph H. and Christopher D. Shaw, **Dynamics—the Geometry of Behavior**, vol. 1 of *The Visual Mathematics Library*, Aerial Press, Santa Cruz, 1981. [ISBN 0-942344-01-4].

(cited on p. 193)

[Altes 1975]

Altes, R.A., "Spline-like Image Analysis with a Complexity Constraint. Similarities to Human Vision," unpublished paper, ca. 1975, 36 pp.

(cited on p. 17,19,35)

[Arnold, R.D. 1983]

Arnold, R. David, "Automated Stereo Perception," Department of Computer Science, Stanford University, Ph.D. thesis, 1983.

(cited on p. 178)

We haven't finished with the regular points, either. The Lie algebra analysis we began can be extended to analyzing the problems of finding the shape and motion of moving objects. The questions of what information is necessary should be resolvable.

We have thus far mainly ignored the problems of photometry. [Koenderink and van Doorn 1980] have pioneered in applying geometric methods here. The Lie algebra approach can be extended to include photometry by considering not just the object in space, but a double sphere bundle over it, describing the directions of light and observer. Part of this is already implicit in the Gaussian sphere approach, for example. There are many interesting results that may be of use. Lines of principal curvature seem to be important, but it is only recently that the topology of the lines of principal curvature, i.e. how they fill out the surface, has been thoroughly understood [Sotomayor 1984].

All this geometry must be brought to bear to get local and global understanding of the image intensity function, the right type of understanding to make deductions about the physical situation that produced it. We have been arguing that an important element is *qualitative*, i.e. geometric understanding, rather than quantitative. A picture is not a C^∞ function, so there is a problem of how to derive this information. The data results from a map from an infinite-dimensional space of smooth functions to a finite-dimensional one of values on a grid, and indeed to a finite set of digitized values. The relation of these maps to the smooth theory has to be looked at carefully. Probably the most direct way to apply theory for smooth functions to this data is to choose some smooth function to represent it, i.e. fit the data. How to do the fit? There are many choices: polynomials, Fourier interpolants, spheroidal harmonics, etc. The mathematics of fitting is partially independent of what is being fit, so it should be possible to obtain a theory without making a choice of basis at the outset. The same philosophy should be transferable to implementation: the program could be designed to take the basis as data.

Postscript

It's customary to conclude a thesis with a compendium of "future research directions," the research that should have been, but wasn't, done for the present work, but will be sometime soon. In adhering to this tradition, I present here a sketch of a program of research that continues what was started here.

The imaging projection has regular points and singular points. Interesting edges occur at the singular points (which are generally limbs), but our geometric analysis has been confined mostly to the regular points, mainly because it's easier. We still had to consider singular points, but they were of lower dimension. A large theory exists for singularities of stable mappings; it is waiting to be applied. [Koenderink and van Doorn 1976, Koenderink and van Doorn 1980, Koenderink and van Doorn 1982] have begun some of *this* work. First you have to classify the singularities which can occur. There are only 2 singularities for generic maps from the plane to the plane: the *fold* and the *cusp*. In a masterly work, [Arnold, V.I. 1983] suggests, however, that the right setting is singularities of a projection from a generic embedding. This is not quite the same as a generic map between surfaces, and Arnold and his coworkers have found that there are exactly 14 types of singularities in this setting.

We have stressed that picture data only reveals geometry via the measuring device of the image intensity. It is therefore necessary to go beyond the projection singularities themselves, and study how they may be inferred from the image function. This is the generalization of the edge detection intuition: "look for discontinuities."

To all this, we can add time. This leads to the study of unfoldings of singularities, and again the time-varying image intensity is the telltale, and the Lie group and Lie algebra of the motion will be the instruments of analysis.

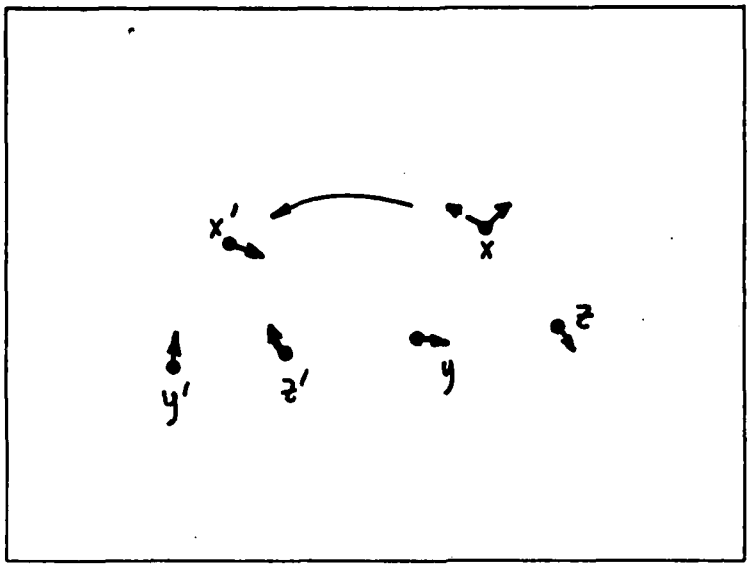


Fig. (6 vector points)

If we remove a vector from one of the 3 original points (i.e. remove a point from the set in $T^*\mathbf{R}^2$), this leaves us with a 1-dimensional kernel in $\mathfrak{e}(3)$. If we go to one of the new points, the spanning lemma tells us we can again measure the kernel, perhaps after an arbitrarily small perturbation. This can be repeated, and 2 more measurements moved the same way, to get 6 vectors at 6 points, corresponding to ∇f, perhaps slightly perturbed.

Now we can see what happens when we choose 6 points in the image. df gives us 6 points in $T^*\mathbf{R}^2$. We can perturb these points to guarantee that $df \not\equiv 0$. Now since every neighborhood of each point maps to a spanning set of $\mathfrak{e}^*(3)$ (local spanning lemma), we can always perturb the nth point so that it is mapped to something outside the span of the first $n-1$ points (at least through $n = 6$, anyway). This gives a perturbation of the 6 points which maps to a spanning set. Since spanning sets are open, these points will still span under sufficiently small perturbation. (In general, one might need a perturbation of both the location of the points and of f to guarantee a spanning set. The degenerate situation occurs when the optic velocity vector is in the direction of constant f.) QED

Here is a more concrete way of looking at the last part of the proof. We already saw that if we had 2 generic functions then we would be finished with 3 generic points. This is the situation of Fig. (3 fibers). It can be pictured in the image as in Fig. (vector points).

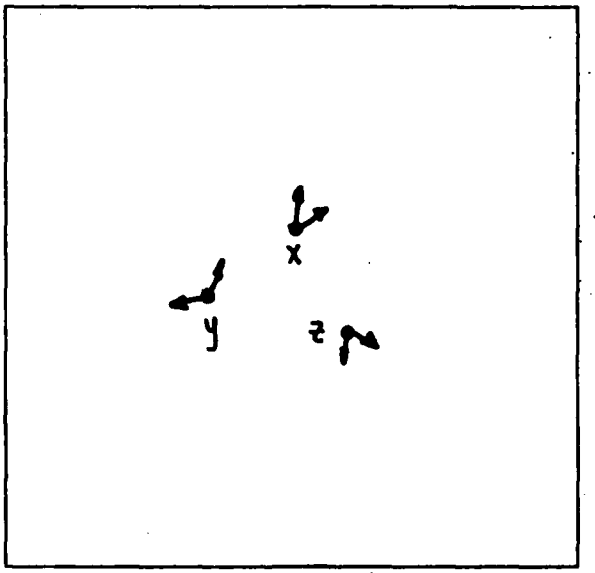


Fig. (3 vector points).

Each vector represents a direction in which the optic velocity vector can be measured, i.e. a value of ∇f for one of the functions. We want to get rid of one of those at each point, and substitute measurements at 3 new points that have been given to us.

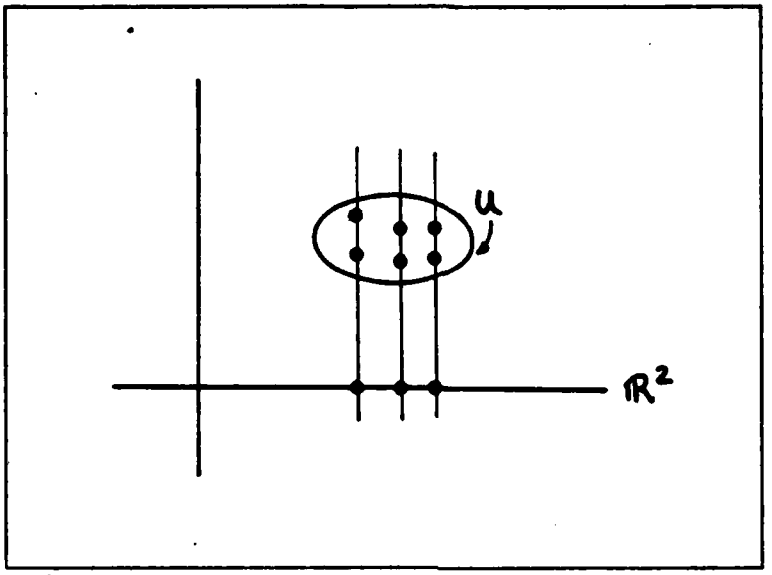


Fig. (local spanning)

Proof. *Choose* a point and neighborhood in $T^*\mathbf{R}^2$. It projects to a neighborhood of $\mathbf{R}^2$, in which we can choose 3 generic points. We can then choose 6 points in $T^*\mathbf{R}^2$, 2 to a fiber, by the 3 fiber lemma. QED (local spanning).

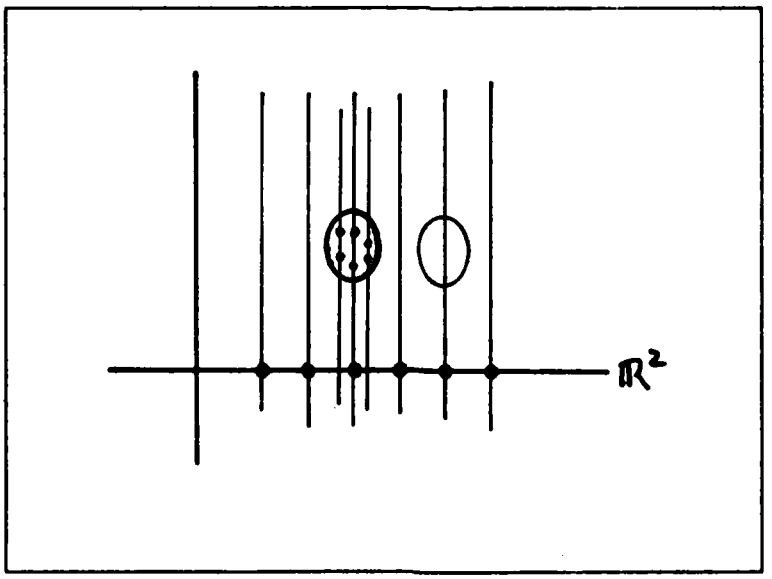


Fig. (6 points)

What we saw earlier is.

Lemma(3 fiber lemma). If we choose 3 generic points in $\mathbf{R}^2$, and 2 linearly independent covectors in each fiber over those points, the 6 resulting points of $T^*\mathbf{R}^2$ are mapped to a spanning set in $\epsilon^*(3)$.

$T^*\mathbf{R}^2$ is 4-dimensional, so it is a little hard to draw. We represent the situation schematically in Fig. (3 fibers).

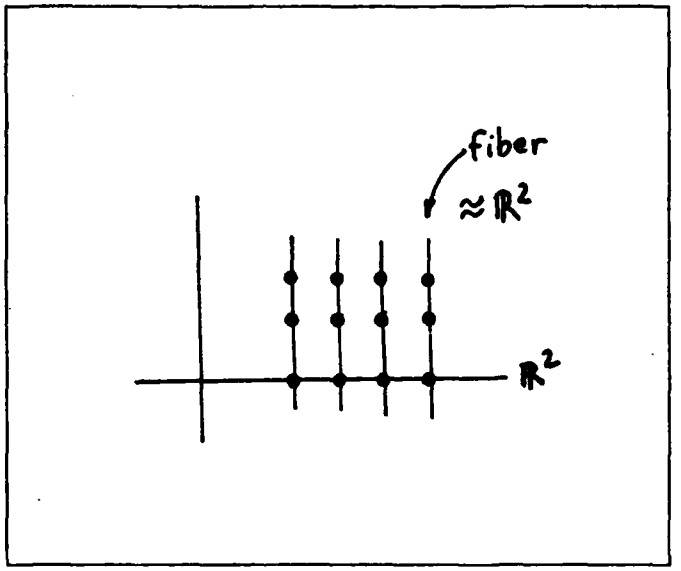


Fig. (3 fibers)

What we will now show is that we can choose *any* 6 generic points in $T^*\mathbf{R}^2$, i.e. 6 generic points in the image, and 6 generic values of df at those points (i.e. a generic f). This is pretty easy by making use of the 3 fiber lemma. The lemma still applies for any neighborhood of $\mathbf{R}^2$, i.e. we can choose the 3 points arbitrarily close together. This gives us

Lemma(local spanning). Every neighborhood of every point in $T^*\mathbf{R}^2$ contains 6 points which are mapped to a spanning set in $\epsilon^*(3)$.

516.36, ISBN 0-387-90073-X (soft), ISBN 0-387-90072-1 (hard), ISBN 3-540-90073-X (soft)].
(cited on p. 162,163)

[Gregory 1977]
Gregory, Richard, "Vision with Isoluminant Colour Contrast: 1. A Projection Technique and Observations," *Perception*, 6, 1977, 113.
(cited on p. 176)

[Griffith 1970]
Griffith, A.K., "Computer Recognition of Prismatic Solids," Ph.D. thesis, MIT MAC-TR-73, 1970.
(cited on p. 24,40)

[Griffith 1973a]
Griffith, A.K., "Mathematical Models for Automatic Line Detection," *Journal of the ACM*, vol. 20, no. 1, January 1973, p. 62.
(cited on p. 24)

[Griffith 1973b]
Griffith, A.K., "Edge detection in simple scenes using a priori information," *IEEE Transactions on Computers*, vol. C-22, April 1973, 371-181.
(cited on p. 24)

[Grimson 1980]
Grimson, William Eric Leifur, **From Images to Surfaces: A Computational Study of the Human Early Visual System**, William Eric Leifur Grimson, MIT Press, 1981. An earlier version was printed as: "Computing Shape Using a Theory of Human Stereo Vision," Department of Mathematics, MIT, Ph.D. thesis, June 1980.
(cited on p. 178,179)

[Guillemin and Pollack 1974]

Guillemin, Victor and A. Pollack, **Differential Topology**, Prentice-Hall, Englewood Cliffs, N.J., 1974. [QA613.5.G84, 514'.7, ISBN 0-13-212605-2].

(cited on p. 157,163)

[Halmos 1957]

Halmos, Paul R., **Introduction to Hilbert Space and the Theory of Spectral Multiplicity**, Chelsea, 1957.

(cited on p. 16)

[Halmos 1963]

Halmos, Paul R., "What Does the Spectral Theorem Say?," *The American Mathematical Monthly*, March 1963, 241–247.

(cited on p. 16)

[Hannah 1974]

Hannah, Marsha Jo, "Computer Matching of Areas in Stereo Images," Ph.D. thesis, Stanford Artificial Intelligence Laboratory, AIM-239, July 1974.

(cited on p. 178)

[Haralick 1980]

Haralick, Robert M., "Edge and Region Analysis for Digital Image Data," *Computer Graphics and Image Processing*, vol. 12, no. 1, January 1980, 60–73.

(cited on p. 19,23,24,29,44,51)

[Haralick 1981]

Haralick, Robert M., "The Digital Edge," *Proceedings of IEEE Conference on Pattern Recognition and Image Processing*, August 1981, 285–291.

(cited on p. 19,53,54,55)

[Haralick 1982]

Haralick, Robert M., "Zero-Crossing of Second Directional Derivative Edge Operator," *Proceedings of SPIE Symposium on Robot Vision*, Arlington, Virginia, May 1982.

(cited on p. 53,54)

[Haralick, Watson, Laffey 1983]

Haralick, Robert M., Layne T. Watson, and Thomas J. Laffey, "The Topographic Primal Sketch," *The International Journal of Robotics Research*, vol. 2, no. 1, spring 1983.

(cited on p. 54,55)

[Haralick 1984]

Haralick, Robert M., "Digital Step Edges from Zero Crossing of Second Directional Derivatives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 1, January 1984, 58–68.

(cited on p. 53,54,55)

[Harlow and Eisenbeis 1973]

Harlow, C. and S. Eisenbeis, "The Analysis of Radiographic Images," *IEEE Transactions on Computers*, vol. C-22, 1973, 678–689.

(cited on p. 25)

[Herskovits and Binford 1970]

Herskovits, Annette and Thomas O. Binford, "On Boundary Detection," MIT Project MAC, Artificial Intelligence Memo 183, July 1970.

(cited on p. 13,24,40)

[Hirsch and Smale 1974]

Hirsch, Morris W. and Stephen Smale, **Differential Equations, Dynamical Systems, and Linear Algebra**, vol. 60 of *Pure and Applied Mathematics*,

Academic Press, New York, 1974. [QA3.P8, QA372 ISBN 0-12-349550-4].
(cited on p. 193)

[Hirsch 1976]

Hirsch, Morris W., **Differential Topology**, Graduate Texts in Mathematics, vol. 33, Springer, New York, 1976.
(cited on p. 157,163,167,173)

[Horn 1972]

Horn, Berthold K.P., "The Binford-Horn Edge Finder," MIT AI Memo 285, 1972, revised December 1973.
(cited on p. 4,20,40)

[Horn and Schunck 1980]

Horn, Berthold K.P. and Brian G. Schunck, "Determining Optical Flow," MIT AI Memo 572, MIT Artificial Intelligence Laboratory, April 1980; also *Artificial Intelligence*, 17, 1981, 185-203.
(cited on p. 206,212)

[Hough 1962]

Hough, P.V.C., "Method and Means for recognizing complex patterns," U.S.Patent 3,069,654, December 18, 1962.
(cited on p. 19,40,74,81)

[Hsu, Mundy, Beaudet 1978]

Hsu, S., J.L. Mundy, P.R. Beaudet, "Web Representation of Image Data," *Proceedings of the Fourth International Joint Conference on Pattern Recognition (IJCPR-78)*, (Kyoto, Japan, November 7-10, 1978), 579-583.
(cited on p. 19,47)

[Hueckel 1969]

Hueckel, M.H., "An Operator which Locates Edges in Digital Pictures," Stanford Computer Science Dept. Memo AIM-105, October 1969.
(cited on p. 17,31)

[Hueckel 1971]

Hueckel, M.H., "An Operator which Locates Edges in Digital Pictures," *JACM*, vol. 18, no. 1, January 1971, 113–125. Erratum in 21, 1974,350.
(cited on p. 17,24,31,38,39,40,41)

[Hueckel 1973]

Hueckel, M.H., "A Local Visual Operator Which Recognizes Edges and Lines," *JACM*, vol. 20, no. 4, October 1973, 634–647.
(cited on p. 24,31,38,39,40,41)

[Hummel and Gidas 1984]

Hummel, Robert A. and Basilis C. Gidas, "Zero Crossings and the Heat Equation," Technical Report no. 111, Robotics Report no. 18, New York University Courant Institute of Mathematical Sciences, March 1984.
(cited on p. 203)

[Iooss and Joseph 1980]

Iooss, Gérard and Daniel D. Joseph, **Elementary Stability and Bifurcation Theory**, Springer-Verlag, New York, 1980. [QA372.I68, ISBN 0-387-90526-X].
(cited on p. 199)

[*Jacobellis v. Ohio* 1964]

Jacobellis v. Ohio, 378 U.S. 184 (1964).
(cited on p. 4)

[Julesz 1960]

Julesz, Bela, "Binocular Depth Perception of Computer-Generated Patterns," *Bell Systems Technical Journal* 39, 1125, 1960.

(cited on p. 176)

[Julesz 1971]

Julesz, Bela, **Foundations of Cyclopean Perception**, Chicago, University of Chicago Press, 1971.

(cited on p. 176)

[Kanade 1978]

Kanade, Takeo, "Region Segmentation: Signal vs. Semantics," *Proceedings of the Fourth International Joint Conference on Pattern Recognition (IJCPR-78)*, (Kyoto, Japan, November 7-10, 1978), 579-583.

(cited on p. 25)

[Kelly 1971]

Kelly, M., "Edge Detection by Computer Using Planning," in **Machine Intelligence VI**, B.Meltzer and D.Michie, eds., American Elsevier, New York, 1971, 397-409.

(cited on p. 25,40)

[Kirsch 1971]

Kirsch, R.A., "Computer Determination of the Constituent Structure of Biological Images," *Computers and Biomedical Research*, vol. 4, no. 3, 1971, 315-328.

(cited on p. 14,21,41,86)

[Koenderink and van Doorn 1976]

Koenderink, J.J. and A.J. van Doorn, "The Singularities of the Visual Mapping,"

Biological Cybernetics, 24, 1976, 51–59.
(cited on p. 148,223)

[Koenderink and van Doorn 1979]
Koenderink, J.J. and A.J. van Doorn, "The Structure of Two-dimensional Scalar Fields with Applications to Vision," *Biological Cybernetics*, 33, 1979, 151–158.
(cited on p. 79,189,193,202)

[Koenderink and van Doorn 1980]
Koenderink, J.J. and A.J. van Doorn, "Photometric Invariants Related to Solid Shape," *Optica Acta*, vol. 27, no. 7, 1980, 981–996.
(cited on p. 51,223,224)

[Koenderink and van Doorn 1982]
Koenderink, J.J. and A.J. van Doorn, "The Shape of Smooth Objects and the Way Contours End," *Perception*, 11, 1982, 129–137.
(cited on p. 148,223)

[Krakauer 1971]
Krakauer, Lawrence J., "Computer Analysis of Visual Properties of Curved Objects," Project MAC, Massachusetts Institute of Technology, MAC TR-82, May 1971.
(cited on p. 189)

[Landau and Pollak 1961]
Landau, H.J. and H.O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — II," *Bell Syst. Tech. J.*, 40, January 1961, 65–84.
(cited on p. 22,56,60)

[Landau and Pollak 1962]
Landau, H.J. and H.O. Pollak, "Prolate Spheroidal Wave Functions, Fourier

Analysis, and Uncertainty — III: The Dimension of the Space of Essentially Time- and Band-Limited Signals," *Bell Syst. Tech. J.*, 41, July 1962, 1295–1336.
(cited on p. 22,56,60)

[Lang 1969]
Lang, S., **Real Analysis**, Addison-Wesley, Reading, Massachusetts, 1969.
(cited on p. 105,106,198)

[Longuet-Higgins 1960]
Longuet-Higgins, M.S., "Reflection and Refraction at a Random Moving Surface. I. Pattern and Paths of Specular Points," *Journal of the Optical Society of America*, vol. 50, no. 9, September 1960, 838–844.
(cited on p. 199)

[Lunscher 1983]
Lunscher, W.H.H.J., "The Asymptotic Optimal Domain Filter for Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 6, November 1983, 678-680.
(cited on p. 56,57,58)

[Marimont 1984]
Marimont, David H., "A Representation for Image Curves," *Proceedings of the 1984 National Conference of the American Association for Artificial Intelligence*, 1984 (to appear).
(cited on p. 126)

[Marr 1982]
Marr, David, **Vision**, W.H. Freeman, San Francisco, 1982. [QP475.M27, ISBN 0-7167-1284-9].
(cited on p. 178)

[Marr and Hildreth 1979]

Marr, David and Ellen Hildreth, "Theory of Edge Detection," AI Memo 518, MIT AI Lab, April 1979. Also Proc.R.Soc.Lond.B., 1980, 207, 187-217.

(cited on p. 18,56,59,61,69)

[Marr and Poggio 1976]

Marr, D. and T. Poggio, "Cooperative Computation of Stereo Disparity," *Science*, vol. 194, October 1976, 283-287.

(cited on p. 178)

[Marr and Poggio 1977]

Marr, D. and T. Poggio, "A Theory of Human Stereo Vision," MIT Artificial Intelligence Memo AIM-451, November 1977.

(cited on p. 178,179)

[Martelli 1972]

Martelli, A., "Edge Detection using Heuristic Search Methods," Dept of EE and Computer Science, NYU, University Heights, Bronx, NY, 10453. Also *Computer Graphics and Image Processing*, 1, 1972, 169-182.

(cited on p. 14,20,24,95,126)

[Martelli 1973]

Martelli, A., "An Application of Heuristic Search Methods to Edge and Contour Detection," Instituto di Elaborazione della Informazione del Consiglio Nazionale delle Richerche, Pisa, 1973. Also *Comm. ACM* 19, 1976, 73-83.

(cited on p. 14,20,95,126)

[Milnor 1965]

Milnor, J., **Topology from the Differential Viewpoint**, Univ. of Virginia Press, 1965.

(cited on p. 164)

[Montanari 1970]

Montanari, Ugo, "On the Optimal Detection of Curves in Noisy Pictures," Artificial Intelligence Laboratory, Stanford University, Memo AIM-115, 1970.
(cited on p. 20,24,93,96,126)

[Montanari 1971]

Montanari, Ugo, "On the Optimal Detection of Curves in Noisy Pictures," *Comm. ACM* 14, May 1971, 335–345.
(cited on p. 20,93,96,126)

[Moravec 1977]

Moravec, Hans P., "Towards Automatic Visual Obstacle Avoidance," *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, MIT, Cambridge, Massachusetts, August 1977, 584.
(cited on p. 178)

[Moravec 1980]

Moravec, Hans P., "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Stanford Artificial Intelligence Laboratory, AIM-340, Ph.D. thesis, September 1980.
(cited on p. 178)

[Morse and Feshbach 1953]

Morse, P.M. and H. Feshbach, **Methods of Theoretical Physics, Part II**, McGraw-Hill, 1953.
(cited on p. 31)

[Murphy 1969]

Murphy, A.S., "An Application of Heuristic Search Procedures to Picture Interpretation," Research Memo MIP-R-61, School of A.I., Edinburgh

University, 1969.
(cited on p. 40)

[Nackman 1982]
Nackman, Lee R., "Two dimensional Critical Point Configuration Graphs," Research Report RC 9642 (#42603), IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., October, 1982.
(cited on p. 193)

[Nackman 1984]
Nackman, Lee R., "Two dimensional Critical Point Configuration Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 4, July 1984, 442–450.
(cited on p. 193)

[Nagel 1983]
Nagel, Hans-Hellmut, "Displacement Vectors Derived from Second-Order Intensity Variations in Image Sequences," *Computer Vision, Graphics, and Image Processing*, vol. 21, 1983, 85-117.
(cited on p. 206)

[Nevatia 1976]
Nevatia, Ramakant, "Depth Measurement by Motion Stereo," *Computer Graphics and Image Processing*, 5, 1976.
(cited on p. 178)

[Nevatia 1977]
Nevatia, Ramakant, "A Color Edge Detector and Its Use in Scene Segmentation," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-7, no. 11, November 1977, 820–826.
(cited on p. 34)

[Nevatia and Babu 1978]

Nevatia, Ramakant and K.R. Babu, "Linear Feature Extraction," *Proceedings of ARPA Image Understanding Workshop*, Pittsburgh, November 1978, 73–78.

(cited on p. 4,20)

[Nitecki 1971]

Nitecki, Zbigniew, **Differentiable Dynamics**, MIT Press, Cambridge, 1971. [ISBN 0-262-14009-8 (hard), ISBN 0-262-64011-2 (paper)].

(cited on p. 162)

[O'Gorman 1976]

O'Gorman, F., "Edge Detection using Walsh Functions," *Proc AISB*, July 1976, p. 195. Also: *Artificial Intelligence* 10, 1978, 215–233.

(cited on p. 17,22,38)

[O'Gorman and Clowes 1976]

O'Gorman, F. and M. Clowes, "Finding Picture Edges Through Collinearity of Feature Points," *IEEE Transactions on Computers*, vol. C-25, no. 4, April 1976.

(cited on p. 40)

[Ohlander 1975]

Ohlander, R.B., "Analysis of Natural Scenes," Dept. of Computer Science, Carnegie-Mellon Univ, April 1975. (PhD thesis).

(cited on p. 21,90)

[Ohta and Kanade 1983]

Ohta, Yuichi and Takeo Kanade, "Stereo by Intra- and Inter-scanline Search Using Dynamic Programming," Computer Science Department, Carnegie-Mellon University, October 1983.

(cited on p. 178)

[Panton 1978]

Panton, Dale J., "A Flexible Approach to Digital Stereo Mapping," *Photogrammetric Engineering and Remote Sensing*, vol. 44, no. 12, December 1978, 1499–1512.

(cited on p. 178,179)

[Pavlidis 1972]

Pavlidis, T., "Segmentation of Pictures and Maps through Functional Approximation," *Computer Graphics and Image Processing*, vol. 1, 1972, 360–372.

(cited on p. 26)

[Peixoto 1973]

Peixoto, M.M., "On the Classification of Flows on 2-Manifolds," in **Dynamical Systems**, M.M. Peixoto (Editor), Academic Press, New York, 1973, 389–419.

(cited on p. 192)

[Pingle 1966]

Pingle, K.K., "A Program to Find Objects in a Picture," Memo AIM-39, Artificial Intelligence Laboratory, Stanford University, January 1966.

(cited on p. 40)

[Pingle and Tenenbaum 1971]

Pingle, K. and J. Tenenbaum, "An Accomodating Edge Follower," Proceedings of the 2nd International Joint Conference on Artificial Intelligence (IJCAI-2), London, September 1971.

(cited on p. 40)

[Poston and Stewart 1978]

Poston,Tim and Ian Stewart, **Catastrophe Theory and its Applications,**

Pitman, London, 1978. [ISBN 0-273-01029-8 (hard), 0-273-08429-1 (soft)].
(cited on p. 199,202)

ızdny 1980]
Prazdny, K., "Egomotion and Relative Depth Map from Optical Flow," *Biological Cybernetics*, vol. 36, 1980, 87–102.
(cited on p. 206)

ızdny 1981]
Prazdny, K., "Determining the Instantaneous Direction of Motion from Optical Flow Generated by a Curvilinearly Moving Observer," *Computer Graphics and Image Processing*, vol. 17, 1981, 238–248.
(cited on p. 206)

ızdny 1983]
Prazdny, K., "On the Information in Optical Flows," *Computer Vision, Graphics, and Image Processing*, vol. 22, 1983, 239–259.
(cited on p. 206)

:witt 1970]
Prewitt, J.M.S., "Object Enhancement and Extraction," in **Picture Processing and Psychopictorics**, B.S.Lipkin and A.Rosenfeld,Eds., Academic Press, New York, 1970.
(cited on p. 19,22,41,53)

am 1971]
Quam, Lynn H., "Computer Comparison of Pictures," Stanford Artificial Intelligence Laboratory, AIM-144, Ph.D. thesis, 1971.
(cited on p. 178)

[Resnikoff 1974]

Resnikoff, H.L., "On the Geometry of Color Perception," in **Some Mathematical Questions in Biology. VI**, S.A. Levin, Ed., *Lectures on Mathematics in the Life Sciences*, vol. 7, American Mathematical Society, Providence, 1974.

(cited on p. 149)

[Roberts 1963]

Roberts, L.G., "Machine Perception of Three-Dimensional Solids," in **Optical and Electro-Optical Information Processing**, J.T.Tippett et al., Eds., MIT Press, Cambridge, Massachusetts, 1965, 159-197. Also Technical Report no. 315, Lincoln Laboratory, MIT (May 1963).

(cited on p. 20,28,40,41,52)

[Robinson 1977]

Robinson, G.S. "Edge Detection by Compass Gradient Masks," *Computer Graphics and Image Processing*, vol. 6, no. 5, October 1977, 492-501.

(cited on p. 41)

[Rosenfeld, Hummel, Zucker 1975]

Rosenfeld, Azriel, R.A. Hummel, and S.W. Zucker, "Scene Labelling by Relaxation Operations," Computer Science Center, Univ of Md, TR-379, May 1975. Also *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-6, no. 6, June 1976, 420-433.

(cited on p. 20,24,96)

[Rosenfeld and Kak 1976]

Rosenfeld, Azriel and A.C. Kak, **Digital Picture Processing**, Academic Press, New York, 1976.

(cited on p. 14)

[Rudin 1964]

Rudin, W., **Principles of Mathematical Analysis**, McGraw-Hill, New York, 1964.

(cited on p. 111)

[Santalo 1976]

Santalo Sors, L.A., "Integral Geometry and Geometric Probability," Addison-Wesley, 1976 (Vol. 1 in Encyclopedia of Mathematics and its Applications).

(cited on p. 81)

[Shafer 1980]

Shafer, S.A., "MOOSE. Users' Manual, Implementation Guide, Evaluation," Bericht 70, IfI-HH-B-70/80, Fachbereich Informatik, Universität Hamburg, April 1980.

(cited on p. 21,91)

[Shanmugam, Dickey, Green 1979]

Shanmugam, K.S., F.M. Dickey, and J.A. Green, "An Optimal Frequency Domain Filter for Edge Detection in Digital Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, January 1979, 39-47.

(cited on p. 15,55,56,58,60,61,63)

[Shapiro 1974]

Shapiro, S.D., "Detection of Lines in Noisy Pictures Using Clustering," *Proceedings of the Second International Joint Conference on Pattern Recognition*, Copenhagen, August 13-15, 1974, 317-318.

(cited on p. 19)

[Shapiro 1975]

Shapiro, S.D., "Transformations for the Computer Detection of Curves in Noisy

Pictures," *Computer Graphics and Image Processing*, 4, 1975, p. 328.
(cited on p. 19)

[Shapiro 1978]

Shapiro, S.D., "Generalization of the Hough Transform for Curve Detection in Noisy Digital Images," *Proceedings of the Fourth International Joint Conference on Pattern Recognition* (IJCPR-78), 710–714.
(cited on p. 19)

[Shaw 1977]

Shaw, G.B., "Local and Regional Edge Detectors: Some Comparisons," Univ. of Maryland Technical Report TR-614, December 1977.
(cited on p. 33)

[Shaw 1979]

Shaw, G.B., "Local and Regional Edge Detectors: Some Comparisons," Computer Graphics and Image Processing, vol. 9, no. 2, February 1979, 135–149.
(cited on p. 33)

[Shirai 1973]

Shirai, Yoshiaki, "A context sensitive line finder for recognition of polyhedra," *Artificial Intelligence*, 4, Summer 1973, 95–119.
(cited on p. 4,25,40)

[Slepian 1965]

Slepian, D., *J. Math. Phys.*, MIT, vol. 44, 1965, 93–99.
(cited on p. 56)

[Slepian and Pollak 1961]

Slepian, D. and H.O. Pollak, "Prolate Spheroidal Wave Functions, Fourier

Analysis, and Uncertainty — I," *Bell Syst. Tech. J.*, 40, January 1961, 43–63.
(cited on p. 22,56,60)

[Smale 1967]

Smale, Stephen, "Differentiable Dynamical Systems," *Bull Amer. Math. Soc.*, 73, 1967, 747–817. Also in Smale, Stephen, **The Mathematics of Time**, Springer-Verlag, New York, 1980. [QA614.S6 ISBN 0-387-90519-7].
(cited on p. 190,193)

[Somerville and Mundy 1976]

Somerville, C. and J.L. Mundy, "One Pass Contouring of Images Through Planar Approximation," *Proc. of the 3rd International Joint Conference on Pattern Recognition (IJCPR-76)*, November 1976 (IEEE 76CH1140-3C).
(cited on p. 21,88)

[Solomayor 1984]

Solomayor, Jorge, "How Lines of Principal Curvature Fill Out a Surface," unpublished, 1984.
(cited on p. 224)

[Streifer 1965a]

Streifer, W., "Asymptotic Expansions for Spheroidal Functions Optical Resonator Modes," *Journal of the Optical Society of America*, vol. 55, July 1965, p. 602.
(cited on p. 57)

[Streifer 1965b]

Streifer, W., "Optical Resonator Modes — Rectangular Resonators of Spherical Curvature," *J.O.S.A.*, vol. 55, July 1965, 868–879.
(cited on p. 57)

[Sunday 1978]

Sunday, Daniel, "The Use of Symmetry Invariants in Feature Extraction Theory," unpublished paper, Dept. of Physiology-Anatomy, Univ. of Calif., Berkeley, 1978.

(cited on p. 119,120)

[Tenenbaum 1970]

Tenenbaum, J.M., "Accommodation in Computer Vision," Memo AIM-134, CS-182, Artificial Intelligence Laboratory, Stanford University, 1970.

(cited on p. 40)

[Thom 1972]

Thom, René, **Stabilité Structurelle et Morphogénèse**, Benjamin, New York, 1972. Translation: **Structural Stability and Morphogenesis**, D.H. Fowler, translator, Benjamin-Addison Wesley, New York, 1975.

(cited on p. 193)

[Tsai 1983a]

Tsai, Roger Y., "3-D Inference from the Motion Parallax of a Conic Arc and a Point in Two Perspective Views," Research Report RC 9818 (#43425), IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.,

(cited on p. 206)

[Tsai 1983b]

Tsai, Roger Y., "Estimating 3-D Motion Parameters and Object Surface Structures from the Image Motion of Conic Arcs," Research Report RC 9819 (#43427), IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.,

(cited on p. 206)

[Tsai and Huang 1984]

Tsai, Roger Y. and Thomas S. Huang, "Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 1, 1984, 13–27.
(cited on p. 206)

[Turner 1974]

Turner, K., "Computer Perception of Curved Objects Using a Television Camera," Ph.D. dissertation, Edinburgh University, November 1974.
(cited on p. 14,40)

[Ullman 1979]

Ullman, Shimon, **The Interpretation of Visual Motion**, MIT Press, Cambridge, 1979. [BF241.U43 ISBN 0-262-21007-X].
(cited on p. 206)

[Waltz 1972]

Waltz, David, "Generating Semantic Descriptions from Drawings of Scenes with Shadows," MIT-AI Technical Report AI-TR-271, 1972. Also "Understanding Line Drawings of Scenes with Shadows," in **The Psychology of Computer Perception**, Patrick Henry Winston, ed., McGraw-Hill, 1975.
(cited on p. 4,97)

[Witkin 1983]

Witkin, Andrew P., "Scale space filtering," *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 1983, 1019–1022.
(cited on p. 79,149,189,202)

[Yuille and Poggio 1983]

Yuille, A.L. and Tomaso Poggio, "Scaling Theorems for Zero-crossings," MIT

Artificial Intelligence Laboratory, AI Memo 722, June 1983.
(cited on p. 203)

[Zucker, Hummel, Rosenfeld 1977]

Zucker, S.W., R.A. Hummel, and A. Rosenfeld, "An Application of Relaxation Labelling to Line and Curve Enhancement," *IEEE Transactions on Computers*, vol. C-26, 1977, 394–403.
(cited on p. 20,96)

END

FILMED

8-85

DTIC